DOCUMENT RESUME

ED 133 363 TM 005 978

AUTHOR	O'Reilly, Robert P.; And Others
TITLE	The Validation and Refinement of Measures of Literal Comprehension in Reading for Use in Policy Research and Classroom Management.
INSTITUTION	New York State Education Dept., Albany. Div. of Research.
PUB DATE	Feb 76
NOTE	424p.; Not available in hard copy due to marginal legibility of tables
EDRS PRICE	MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS	Class Management; *Cloze Procedure; Criterion Referenced Tests; Elementary Secondary Education; Item Analysis; *Multiple Choice Tests; Productivity; *Reading Comprehension; Reading Programs; *Reading Tests; Standardized Tests; Test Construction; Testing Problems; Test Reliability; *Test Validity
IDENTIFIERS	Domain Referenced Tests; *Literal Comprehension; Rasch Model; SPPED; SPPED Test Development Notebook; System Pupil Program Evaluation Development

ABSTRACT

The report proposes to complete the validation and refinement of a new domain referenced testing technology designed to assess literal comprehension ability in students in grades 1-12. The domain referenced measures in this technology, along with other more traditional measures of reading comprehension, literal and non-literal, are subsequently intended to be used in part in large scale studies of productivity in school reading programs. To date, studies of productivity in reading instruction have had little influence on educational decision-making due to serious methodological problems, one of the major problems being the lack of adequate measures of program output. The report further proposes to solve a number of important instructional management problems created by the use of the inadequate information available from traditional measures of reading comprehension. The new domain referenced measures of reading comprehension will have an improved basis for scaling students on comprehension ability, and abilty scores from this scale will be referenced to an additional scale defining an individual or group's ability to read in several domains of written discourse. These scaling features will allow for the assignment of students to specific levels of reading materials in specific instructional or content domains, a procedure not possible with existing measures of reading comprehension. (Author)

Documents acquired by ERIC include many informal unpublished materials not available from other sources. ERIC makes every effort to obtain the best copy available. Nevertheless, items of marginal reproducibility are often encountered and this affects the quality of the microfiche and hardcopy reproductions ERIC makes available via the ERIC Document Reproduction Service (EDRS). EDRS is not responsible for the quality of the original document. Reproductions supplied by EDRS are the best that can be made from the original.

ED133363

The Validation and Refinement of Measures of Literal Comprehension in Reading for Use in Policy Research and Classroom Management

Robert P. O'Reilly, R.T. Schuder, Steven J. Kidder,
Ruth Salter, Paul D. Hayford

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

The University of the State of New York
The State Education Department
Division of Research
Albany, New York 12234

February 1976

TM005 978

ABSTRACT

The report proposes to complete the validation and refinement of a new domain-referenced testing technology designed to assess literal comprehension ability in students in grades 1-12. The domain-referenced measures in this technology, along with other more traditional measures of reading comprehension, literal and non-literal, are subsequently intended to be used in part in large scale studies of productivity in school reading programs. To date, studies of productivity in reading instruction have had little influence on educational decision-making due to serious methodological problems, one of the major problems being the lack of adequate measures of program output.

The report further proposes to solve a number of important instructional management problems created by the use of the inadequate information available from traditional measures of reading comprehension. The new domain-referenced measures of reading comprehension will have an improved basis for scaling students on comprehension ability, and ability scores from this scale will be referenced to an additional scale defining an individual or group's ability to read in several domains of written discourse. These scaling features will allow for the assignment of students to specific levels of reading materials in specific instructional or content domains, a procedure not possible with existing measures of reading comprehension.

STATUS OF THE PROJECT:

The proposed work has evolved through several years of research and development on major issues relating to the assessment of school achievement. Prior efforts relating to the present work include the preparation of a bank of instructional objectives defining reading performance, the

development and validation of a criterion-referenced evaluation system known as Comprehensive Achievement Monitoring, and experimentation with the measurement of resource utilization in reading programs as an initial attempt to improve the methodology for productivity research.

Most recently, this research has turned to the development of more adequate measures of reading outcomes--a major gap remaining in productivity methodology. The intent is to produce a test development resource that will be useful at a variety of institutional levels and a measure that will be unique in at least two major respects: (1) it will be a measure of both comprehension achievement and ability, and (2) it will be the only extant and broadly applicable measure of literal comprehension as such--the important, generalized reading skill that underlies all higher-order reading comprehension abilities. Some two years of development effort have culminated in new measures of reading comprehension that are referenced to several major domains of reading materials relevant to students in grades 1-12. These measures are the components of a flexible test-assembly device referred to as the Test Development Notebook or TDN. The TDN, as currently conceived, is a resource for the assembly of measures of literal comprehension in grades 1-12, across all major content domains relevant to the school population.

The content of the TDN developed to this point consists of the multiple-choice cloze component and an alternate measure of the construct of literal comprehension based on the wh-item. The multiple-choice cloze component (referred to as the MCC) consists of approximately 1,500 clozed passages (generally, 60-70 word passages with ten deletions and accompanying multiple-choice items) categorized (temporarily) by readability levels determined by Spache and Dale-Chall readability formulas. The wh-/main idea item pool consists of 300 passages, 15 at each of 20 readability levels. Passage length varies systematically by readability level (e.g., approximately 25 words at level 1 and up to 220 words at levels 17-20). Each of these passages is accompanied by as many as four multiple-choice main idea items and up to eight multiple-choice wh-detail items modeled after Bormuth's (1970) wh-items. The formats of the cloze and wh-materials are both objective, generative procedures for preparing numbers of parallel, multiple-choice items.

The first field test of the MCC and wh-item tests was conducted in May 1975, in an administration of both types of tests in a survey design to approximately 5,000 students spread more or less evenly over grades 1-9. This administration fulfilled several purposes: (1) It explored the use of the testing materials in applying a survey design in one textual area; (2) it provided data for detailed item analyses; (3) it provided an initial test of the ability of the system in assembling large numbers of parallel test forms; (4) it provided a basis for testing out the Rasch or latent-trait model as an approach to scaling; (5) it made available reliability data on a large number of test forms; and (6) it provided initial convergent and discriminant evidence on the validity of the construct. The more important conclusions that were drawn from the field test are as follows:

1. The existing paper-based model of the TDN allowed assembly of 36, 50-item MCC test forms in a matter of a few hours.

2. The application of the survey design model in grades 1-9 was generally successful, for both the MCC and wh-item tests, but the design can be improved in the future by raising the ceiling of readability for upper-grade test batteries.

3. The item analysis data showed that the MCC item format, as applied to a given reading passage, generally yielded a set of items that were consistent and homogeneous within and between passages, regardless of passage level. (The data provided many important leads as to how the homogeneity of items within passages might be improved, but, in general, extensive improvements were not required.)

4. Large numbers of virtually parallel tests could be systematically assembled from the TDN from either the MCC or wh-item compoments. With improved scaling, the possibility of objectively assembling _n_ tests with specified properties is assured, thus providing for transferability of test generation.

5. The experimental application of the Rasch model to 216 MCC test passages showed that the ratio scale properties of this model could be achieved with the item form.

6. Analyses of the reliability of the MCC test forms showed that the tests assembled for the study were highly precise across all grade levels in the study sample. The level of precision is sufficiently high to warrant use of the tests at the individual level. The reliability data further support the inference that the MCC test is reliable over short intervals (i.e., alternate forms of the same test will scale individuals similarly on test-retest with a high degree of precision). The reliability characteristics of the wh-item tests were similar to those achieved with the MCC test.

7. There were several indications of support of the construct validity of the cloze test in the data analysis. The internal consistency measures and the Rasch analyses indicated the MCC test could be accurately described as measuring a homogeneous trait across grades 1-9. The validity coefficients between the MCC test and the wh-item test, an alternate measure of the construct, were consistently high (_r_ = .81 at grades 1-3), except where attenuated by range of talent. The MCC test generally correlated at appropriate levels with measures of verbal and non-verbal IQ, California Achievement Test (CAT) subscores in language and reading, and a measure of passage dependency. The MCC and wh-item tests converged in having virtually identical correlations with the CAT subscores, the IQ scores, and the score on passage dependency. Overall, the results were highly consistent across the 9 grade levels, lending considerable credibility to the validity of the MCC test.

8. The analyses of these field-test data continue to date, as well as use of the data to refine the MCC corpus of passages. Of particular interest is a factor analysis of the test data to be run shortly.

PROPOSED RESEARCH AND DEVELOPMENT

The data analyses on the reliability and validity of the MCC and wh-item formats continue to date. More detailed results, including factor analyses, will be reported in a series of papers at the annual conference of the American Educational Research Association and the National Council on Measurement in Education this spring. The overall results to date, together with reviews by a panel of well-known professionals in reading, psycholinguistics, and educational measurement, have amply demonstrated the desirability of completing the proposed work on the testing materials.

The proposed work on the testing materials is designed to bring the TDN to a state where it can be used as a valid assessment device in a variety of evaluation contexts at state and local levels. The research effort will continue the study of the reliability and content validity of the testing materials, but will focus largely on construct validation, scaling, and packaging.

Construct Validation

The proposed approach for further validation and refinement of the testing materials is a series of concurrent efforts designed both to study the meaning of the tests and to bring them to a broadly usable state. A set of preliminary studies will focus on further refining the MCC test format (the measure of major interest) in preparation for a cross-sectional, longitudinal study of test validity in a sample of approximately 13,000 students in grades 1-12.

The preliminary validity studies will generally determine the boundaries of written discourse to which an MCC test score can be expected to generalize (i.e., Does the meaning of the test score change when passages vary extensively in terms of syntactic and semantic complexity or content area?). In addition, specific features of the item format and the conditions of test administration will be studied to determine any additional refinements that might be made to the test.

The major effort of the proposed validation--the cross-sectional, longitudinal study--will examine the boundaries of the construct of literal comprehension in an expanded matrix of different textual, psycholinguistic, situational, and psycholological factors. The longitudinal study will be conducted in a single urban school district that will contribute a heterogeneous sample of more than 1,000 students in each grade from 1 through 12. The design of the study will provide a developmental context within which the contributions of important school and non-school factors to the MCC test, the wh-item test, and other measures of reading comprehension can be studied across the 12 years of public schooling. The extent to which the various measures of reading comprehension change across the years of schooling can be estimated through this design as well as the proportion of test score change that is attributable to manipulable factors, such as reading experiences in the home or school. Since standardized measures of reading comprehension will also be available for grades 1-9 of the study population, the design will enable a direct and critical comparison of the sensitivity of the various comprehension measures in accounting for the influences of instruction and related experience.

Scaling

Completed work on applying the Rasch model to the TDN passage and item corpus supports the present proposal to calibrate all such passages on a single underlying scale with ratio properties. This application of scaling involves mounting a complex linking design in which both the MCC and wh-item pools will be calibrated using a sample of approximately 50,000 students in grades 1-12. The proposed design will result in the calibration of all test passages in the various content domains covered by both tests on a common Rasch scale. Then all of the many tests that can be assembled from the MCC content domains will be referenced to the same scale.

The proposed major calibration of the test passages will be preceded by a pilot study in which the complexities of the linking design will be worked out by experimental application of the Rasch model to the MCC passages in several content areas outside the basal reader area. The proposed research on scaling further includes the construction of derived scores for the MCC test and the establishment of formal procedures for linking Rasch ability scores with the distributions of readability in related domains of materials.

EXPECTED CONTRIBUTIONS:

The project is expected to make a number of theoretical and practical contributions to improved evaluation in reading and ultimately to improved instruction and better resource allocation at several levels of the educational enterprise. Concurrent with the validity studies proposed for the testing materials, a program will be mounted to gradually transform the TDN into a state of broad practical utility. The principle elements of this program include computerization of the processes of test item generation and test assembly (the former process applicable to the cloze format only) and the preparation of textual materials presenting simulations and guidelines for application of the testing materials in a variety of evaluation contexts. The specific products expected from this and other components of the proposed research and development are:

1. A testing package (the TDN) with a finalized version of the multiple-choice cloze and wh-item testing materials along with a handbook and training materials for its use.

2. A technical report on the readability and other characteristics of reading materials in the domains covered by the testing materials.

3. A technical report on the use of the testing materials in a pilot productivity study.

4. A report or book on the validity of major norm-referenced tests of reading comprehension (SAT, CAT, ITBS, etc.) from the point of view of theory and content.

5. Periodic and final reports on the activities conducted and the results obtained during the funding period.

PROJECT MANAGEMENT:

The research and development proposed here will be conducted by the Bureau of School and Cultural Research, a unit that has six years of experience in the development of criterion-referenced testing in both reading and mathematics. With the aid of nationally known consultants in certain highly specialized areas, such as scaling and decision theory, the Bureau will assemble a technically and professionally competent staff for the proposed task.

The objectivity and technical adequacy of the Bureau's proposed and completed work on the task will be maintained by periodic external review by a panel of nationally-known experts in such fields as psycholinguistics, cognitive development, reading theory, psychometrics, computer technology, and statistics. The required technical facilities for completing the task exist in the Education Department.

Acknowledgments

The authors gratefully acknowledge the contributions to this report of Ronald Streeter and Gerlach van Gendt, Assistants in Education Research, and Joan Heffler, Consultant. The technical expertise of Mr. Streeter and Mr. van Gendt yielded the statistical analyses presented in the report, and the efficiency of Mrs. Heffler facilitated presentation of results and preparation of the text.

The authors also acknowledge with thanks the steadfast support of the clerical and secretarial staff of the Bureau of School and Cultural Research: Kim DeLoria, Kathleen Mattice, Vicki Noble, Debra Tate, and Nancy Wait.

TABLE OF CONTENTS

Page

ABSTRACT ii

ACKNOWLEDGMENTS. viii

INTRODUCTION x

Chapter

I. STANDARDIZED NORM-REFERENCED TESTS OF COMPREHENSION 1-1

II. COMPREHENSION 2-1

III. THE CLOZE PROCEDURE 3-1

IV. SPPED CLOZE EXERCISES IN A MULTIPLE-CHOICE FORMAT 4-1

V. APPLICATION OF THE MULTIPLE-CHOICE CLOZE IN MEASUREMENT AND EVALUATION. 5-1

VI. TEST VALIDATION AND REFINEMENT. 6-1

VII. CALIBRATION AND SCALING DESIGN. 7-1

VIII. ANALYSIS OF CLOZE PASSAGES AND ITEMS. 8-1

IX. RELIABILITY AND VALIDITY OF THE MULTIPLE-CHOICE CLOZE AND WH-ITEM TESTS 9-1

REFERENCES R-1

APPENDICES

A. Procedures for Development of Multiple-Choice Cloze and Wh-Item Components of the TDN A-1

B. Analytical Tables of Multiple-Choice Cloze Items and Passages. B-1

INTRODUCTION

Dismal reports on the level of literacy in American schools and colleges--and in the nation as a whole--appear with relentless regularity in magazines and newspapers. Statements on alarmingly high levels of "functional illiteracy" and declines in student reading achievement abound in spite of efforts to upgrade reading performance through massive expenditures on ESEA Title I programs, the Right to Read, and other special projects. With this contradiction of increased effort and diminishing returns, questions may be raised as to the bases on which judgements are made. What is meant by "literacy," and how is the achievement of specifiable levels of literacy measured?

Assuming that "literacy" refers to minimal competence in reading comprehension, what is lacking is a valid, accurate measure of literacy, a means of determining minimum competency in reading comprehension. If so, those who are concerned with the state of reading in America--and with productivity in American schools--should first be concerned with the availability of appropriate measures of literacy-related outcomes of school reading programs.

This report is concerned with just such a measure. Its focus is the development of an accurate, useful, and economical test of literal comprehension, a fundamental reading skill and the skill involved in what is usually meant by functional literacy. The particular innovative measure, the subject of the report, is the SPPED multiple-choice cloze format.

The first four chapters of this report present a theoretical rationale for the SPPED multiple-choice cloze. They contain a critique of traditional measures of reading comprehension; a discussion of psycholinguistic theory relative to efforts to measure comprehension; a brief discussion of the conventional cloze procedure as a test of comprehension; and a statement of a tentative construct of literal comprehension, including one of its operationalizations in the multiple-choice cloze format developed for the SPPED Test Development Notebook. The fifth chapter describes the advantageous properties of the SPPED multiple-choice cloze which should make it a broadly useful as well as critically important tool for measurement and evaluation.

An overview of the research to date and of the future research and development planned for the multiple-choice cloze and related materials is presented in Chapter VI. Chapter VII outlines a detailed plan for calibrating the multiple-choice cloze passages on a ratio scale based on application of the Rasch model to the cloze testing materials. Together, these plans are designed to bring the cloze testing materials to a broadly usable state in policy research on reading and in the management of reading instruction.

The eighth chapter provides a detailed discussion of both conventional and Rasch item-analysis data available from a preliminary administration of the multiple-choice cloze testing materials to a sample of 5,000 students in grades 1-9. These item analysis data and critical examination of the testing materials show that departures from the expected characteristics of the testing materials are infrequent. Further, current review of all extant multiple-choice cloze materials promises to diminish an already low incidence of errors in procedures and execution.

The ninth chapter, as well as part of the seventh, reports research to date which suggests that the preliminary test development and administration of the SPPED multiple-choice cloze has been highly successful. The cloze testing materials and an alternate measure of literal comprehension developed for the research, called the wh-item test, were shown to be highly internally consistent in the study population. The preliminary research data further provided substantial indications of the validity of the cloze format. The overall results of the research show that the hypothetical advantages proposed for the multiple-choice cloze format (e.g., objectivity, economy, flexibility, domain-referencing) can be broadly realized.

CHAPTER I

STANDARDIZED NORM-REFERENCED TESTS OF COMPREHENSION

As noted in the Introduction to this report, the American people will always hold their school systems accountable for the literacy of students, yet teachers have been provided with neither an acceptable standard of literacy nor the tools to measure the basic reading abilities implied by "literacy." Standardized, norm-referenced tests of reading comprehension, in spite of their many disadvantages, certainly have a place in educational testing,[1] but they are entirely inadequate as measures of ability or achievement in literal comprehension. In the first place, the standardized, norm-referenced tests used to measure achievement in reading comprehension are patently biased toward a conceptualization of reading as reasoning. Besides this, they present at least four additional problems: (1) They are too subjective in construction to be reproduced or properly validated; (2) their scaling properties make scores difficult to interpret; (3) they are insensitive to individual gain or growth; and (4) they are rigid in format and therefore limited in utility.

Reading as Reasoning

Test-makers seldom specify the conceptualization of comprehension behind their tests, much less the psycholinguistic theory and experimental evidence supporting such a conceptualization. Finding no explicit constructs, other than labels on subsections of the tests, researchers are forced to use various analytical techniques to tease out the notion of comprehension from test scores.[2]

---

[1] Primarily as predictors of academic success (Anderson, 1972; Carver, 1974; Thorndike, 1973-74).

However, it is not inappropriate for the consumer of such tests to ask why there is no explicit statement of what it is the test attempts to measure and, given no explanation, to speculate on the reason for its absence. Messick (1975), for instance, notes a long-standing bias against construct validation in educational measurement. Test-makers seem to assume that "educational measurement is primarily concerned with what a pupil can do, and [that] the nature of the accomplishment is clear from the specification of the tasks" (p. 958). But the very terms used to label tests and to interpret test scores "imply process interpretations, such as scientific reasoning or reading comprehension..." (p. 958). Popham (1975) argues (less generously) that the commercial publishers who create and market standardized tests "are loath, from a marketing viewpoint, to spell out exactly what their exams measure" because the tests must be marketed nationally, and "many educators would find them inconsistent with local instructional programs" (sec. 2, p. 4). In any event, without an explicit statement of theoretical and empirical relationships, "the burden of construct validation [is foisted] onto the consumer, who will inevitably make inferences beyond the universe of situations representatively sampled by the test" (Cronbach, 1971, p. 483). Instead of stating an explicit construct, which is subject to rival interpretations, the publishers of standardized comprehension tests usually present the consumer with correlations between standardized tests of the same ilk. But correlations between equally ambiguous tests are small solace to consumers who would like to know what any of them actually measure.

---

[2] For a recent, critical review of "psychometric research on comprehension in reading," see Davis (1972).

While standardized comprehension tests are notorious for their lack of explicity stated constructs, most seem, in fact, to be based on Thorndike's (1917) introspections on reading. In his conceptualization of comprehension, Thorndike made no distinction between reading and thinking:

> Understanding a . . . printed paragraph is then a matter of habits, connections, mental bonds, but these have to be selected from so many others, and given weights so delicately, and used together in so elaborate an organization that "to read" means "to think" as truly as does "to evaluate" or "to demonstrate" or "to verify." (p. 114)

Not only did Thorndike find reading and thinking conceptually indistinguishable, but the extension of the comparison to evaluation, demonstration (proof), and verification implies the equation of reading and "high order" thinking or reasoning processes:

> The successful response to a question or to a paragraph's meaning implies the restraint of tendencies of many words to be over-potent and the special weighting of other tendencies. This task is quite beyond the power of weak minds and is of the same selective and coordinating nature as the more obvious forms of reasoning in mathematics or science. (p. 114)

Thorndike's conceptualization of reading as a thinking or reasoning process has had enormous influence on the teaching of reading (Davis, 1972) and the construction of comprehension tests.

Now few people will deny that comprehension involves thinking (processing information) or that critical/evaluative reading and reasoning share some intellectual skills (e.g., deductive and inductive reasoning). But tests that overemphasize critical/evaluative reading skills at the expense of more fundamental skills like those of literal comprehension, for example, have a limited utility. Teachers, after all, have long felt it necessary to distinguish between "reading the lines, reading between the lines, and reading beyond the lines" when teaching such a complex behavior as reading comprehension. They are aware that, as Feder noted in 1938, "the tasks of answering factual questions and of making inferences call to a considerable extent on quite different fundamental skills in comprehension" (Davis, 1972, p. 658.) The movement toward teaching by objectives and mastery learning has made such distinctions even more important. Tests that stress reasoning processes fail to give proper emphasis to basic comprehension skills that are developmentally and logically prior to more extensive processing of information in a text.

Other than a few token items labeled "literal comprehension," traditional tests of reading comprehension, following Thorndike, make no such distinctions. They are so biased toward a conceptualization of reading as reasoning that they hardly constitute tests of comprehension as such. Beyond fourth grade, when reading instruction concentrates on comprehension, items on reading comprehension tests become increasingly indistinguishable from verbal items on IQ tests (Singer, 1973). Besides correlating with IQ tests of general verbal ability, traditional comprehension tests even correlate substantially with non-verbal, figure-analogies tests of intelligence (Carroll, 1972). Obviously, a student needs some modicum of intelligence, especially in symbolic processes, to be able to read at all, but if the "acquisition of symbol-sound correspondence is within the mental range of a group of students and instructional conditions allow adequate time for achieving the task, then IQ may have a significant relationship to rate of acquisition but not to accomplishment of the task" (Singer, 1973, p. 1).

Passage Dependency. Traditional comprehension test items are so biased toward reading as reasoning that students can score well above chance on significant numbers of test items without bothering to read the passages upon which the questions are supposedly based. And yet any reading comprehension test purports "to measure how well a student understands what he is reading. The questions used to ascertain the degree of this understanding are based on the tacit assumption that a direct relationship exists between reading a passage and answering questions about it" (Tuinman, 1973, p. 208). Weaver, Bickley, and Ford (1969) tested that assumption with samples from many standardized tests of reading comprehension. They discovered that college students who did not read the passages upon which the questions were based answered 67% as many questions correctly as college students who did read the passages. Obviously, many questions were not passage-dependent. The passages, that is, were not the only sources of the information needed to answer the questions. A more recent study of passage-dependency by Tuinman (1973) in grades 4, 5, and 6 found that the "average probabilities of correct responses with no passage present ranged between .32 and .50--well above the expected chance score of .25" (p. 206). The norm-referenced tests used in this study were (a) The Nelson Reading Test, (b) The California Achievement Test, (c) The SRA-Achievement Series, (d) The Metropolitan Achievement Test, and (e) The Iowa Test of Basic Skills.

Processing information derived from written text may well be similar to processing information derived from other verbal and non-verbal sources (Smith, 1975). But when standardized reading comprehension tests stress inferential and related reasoning processes to the extent that the information in the text becomes superfluous to the test items, then the conceptualization of reading comprehension implied by such tests strains credibility. Rather than making inferences about what the text means as a consequence of having read the text, students can infer "the meaning" of the text (as interpreted by the test-writer) from the test items themselves. The information that is assimilated to cognitive structure may not be derived from the text. Therefore, scores on such tests cannot be used as evidence that the students did in fact comprehend "the text." These scores imply comprehension of the test items rather than comprehension of the text itself.

Besides straining the analogy between reading and reasoning, the passage independence of the items on standardized reading comprehension tests raises more serious questions about the objectivity, utility, and validity of such tests. Most instructional reading programs, including those that teach "sampling procedures," necessarily promote a careful perusal of the text. Indeed, how are disputes about the meaning of a text ever resolved except by reference back to the text itself? (The relevancy of biographical and other extra-textual information, for instance, can only be determined by reference to the meanings implied by the text itself.) What use is a reading teacher to make of scores from "reading comprehension tests" that invite students to ignore the text, that promote "comprehension" skills specific to test-taking rather than comprehension skills in general? In fact, teachers characteristically develop comprehension skills by using questions to direct attention to salient features of the text, and, in doing so, they run the risk of training students to validate the teacher's interpretation of the text at the expense of the students' own perceptions. But teachers have a saving grace; they are in a position to recognize and promote the student's independent efforts to interpret the text. No such opportunities exist on tests. Given the multiplicity of interpretations to which most segments of connected discourse are subject, what justification is there for the idiosyncratic interpretations represented by the questions and "correct" answers on any given standardized reading comprehension test? Granted, most skillful readers would usually accept the validity of the interpretation of the text implied by most of the test items on standardized tests. But why any particular interpretation at the arbitrary exclusion of others in a test which claims to measure the general ability to apprehend the meaning of printed discourse? Or do the test items represent a random sample of all possible interpretations? Clearly not. No two test writers interpret a given text in the same way, and this again raises the problem of specifying what standardized comprehension tests actually measure.

Factor Analyses. Factor analyses of scores from standardized comprehension tests, rather than clarifying what such tests measure, only reveal the hodgepodge conceptualizations underlying them. Vocabulary knowledge, test-taking skills, and comprehension skills are all subsumed under a vague, global notion of "comprehension." Davis (1941), for instance, first identified several hundred "reading comprehension skills," and then, noting a considerable overlap, reduced them to nine "test-able skills" (1944). In 1968, he reaffirmed the independent existence of eight of these skills.

Davis'eight unique skills are listed in Table 1.1. Of these, Skill 3--finding answers to questions, with a significant 13 and 7 percent of nonchance variance--can be excluded because it is a test-taking rather than a comprehension skill. (The test items themselves introduce reasoning and inferential processes and difficulties which may be extraneous to the actual comprehension process [Bormuth, 1970]). Moreover, Carroll (1972), noting "the unique variance residing in the tests of these skills," "is tempted to conclude that perhaps only four or five of them merit recognition as distinct skills, and even these are rather highly correlated in high-school populations" (p. 2). Excluding Skill 3, the remaining skills of

Table 1.1

Per cent of Nonchance Variance of Each of Eight Skills That Is Unique in the Set of Skills Used[a]

(N = 988)

Skill	Cross validation by Items and day [b]	Items only
1. Recalling word meanings	35	29
2. Drawing inferences about the meaning of a word from context	- 1	8
3. Finding answers to questions answered explicitly or merely in paraphrase	13	7
4. Weaving together ideas in the content	5	5
5. Drawing inferences from the content	23	18
6. Recognizing a writer's purpose, attitude, tone, and mood	14	8
7. Identifying a writer's techniques	8	3
8. Following the structure of a passage	15	12

Note: From "Research in Comprehension in Reading" by F. B. Davis, Reading Research Quarterly, 1968, 4, 499-545.

[a] The negative entry in the table probably represents a chance deviation from a zero or slightly positive true value.

[[b] An equivalent form of the same test was given to the same students after an interval of one or two days.]

significance--recalling word meanings; drawing inferences from the content; recognizing a writer's purpose, etc.; and following the structure of a passage--represent polar extremes in a hierarchy of reading skills, as would be expected from tests which seem so indebted to Thorndike's conceptualization of comprehension.

The largest nonchance variance is represented by the skill of recalling word meanings. But the skill is "measured by recognition vocabulary items"

(Davis, 1972, p. 663)--words in isolation, that is, or words in limited contexts (usually no more than a phrase). Now obviously word knowledge is necessary to comprehension, but the skills involved in recognizing words in isolation or in limited contexts are quite different from the skills involved in interpreting the interrelationships between word meanings and syntax in connected discourse. Skill 2, on the other hand, deals with words in context and is much closer to comprehension of connected discourse, but it represents only an insignificant percentage of nonchance variance in traditional comprehension tests.

While the major variant, "recalling word meanings," "as measured by recognition vocabulary items," seems to lie outside the pale of reading comprehension (apprehending the meaning of connected discourse), the remaining skills--drawing inferences, recognizing purpose, etc., and following structure--represent the upper reaches of the hierarchy of comprehension skills, the reasoning processes. From his analyses, Davis drew the general conclusion that . . . [comprehension] is largely dependent on knowledge of word meanings and on ability to reason in verbal terms" (Davis, 1972, p. 663).

Subjectivity

To be maximally useful in measuring achievement in reading comprehension, a test must be objective enough to be reproducible. That is, several test writers working independently with the same corpus of materials must be able to produce essentially the same test. What this means in practice is that test writers, when selecting the materials to be included in the test and writing questions about those materials, must follow a detailed, explicit rules system (somewhat like a computer algorithm) which radically limits the opportunity to make subjective decisions based on personal biases and idiosyncracies. Several advantages are gained by such objectivity: (a) if the test is reproducible, there is an objective basis for claiming that two different forms of the test should have the same label (e.g., "reading comprehension"); (b) it becomes possible to examine otherwise arbitrary claims about what the test actually measures, for its genesis is public and traceable; (c) it also becomes possible to compare the results of two different tests in relation to the reading skills being measured; and (d) different forms of the same test can be compiled easily and used to monitor reading development over short periods of time.

Unfortunately, test development procedures for standardized tests of reading comprehension fall far short of this kind of objectivity. Publishers have developed a careful, traditional procedure for constructing standardized tests, but subjectivity is apparent at every stage of the process. Test writers begin, for instance, by developing an outline of the information the test will cover. But since "the outlining procedure is ill-defined, it is difficult to verify that an item measures the content claimed by the label" (Bormuth, 1970, p.12). Then, the passage-sampling procedure is not objective. Once the passages are selected, the test writer is constantly making subjective decisions about which questions to write on each passage. Some questions are rejected as too easy; others as too difficult or too wordy, and so on. The result, as Bormuth (1970) has commented, is that the test writer is "implicitly designing the test" as he goes along, "but doing so in a manner that is not open to inspection and . . . review" (p.13). Perhaps it is precisely the relevant course content that is present in the final form of the test, but the substantial lack of objectivity makes verification impossible.

Scaling Properties

To be precise in measuring gain or growth, an achievement test scale must have equal intervals and a meaningful zero point. A ruler, for instance, is a measurement device with equal intervals and an absolute zero. An inch at either end of a ruler is still an inch, or an inch in linear space is equal to any other measure of one inch in linear space. But part of the "meaning" of that measure of one inch is the possibility of zero length or no inches. The interval of one inch is an absolute measure that does not need to be transformed for comparison with another measure in inches. Once a test is developed to measure gain on a scale with equal intervals and a meaningful zero point, it becomes possible to interpret differences in raw scores as true quantitative measures of gain or growth within individual students over a period of time.

In addition to equal interval scaling and a meaningful zero point, a useful test development procedure must be based on person-free item calibration and item-free person measurement. Such a procedure would result in test scores that could be interpreted in terms of an absolute scale (person-free) rather than in relation to the particular students who took part in the original calibration of the test. The procedure would also produce test scores that would not be dependent on the particular items used on the test (item-free). Reading comprehension tests scaled in this way would result in measures of achievement on a scale from "little ability" to "maximum ability." Interpretations of raw scores would be referenced directly to this equal-interval, meaningful-zero scale. Equivalent and parallel test forms could then be assembled for accurate, periodic testing. School districts could also compare the effects of different educational treatments on individual students or groups of students.

But norm-referenced tests do not have these scaling properties. Instead, test scores are referenced to the particular group of students used to norm the test. The estimate of reading ability these tests produce is dependent upon particular people and the specific content of the test. Comparisons between test scores on different forms of the test are made difficult, in part, because the content of the two forms is not necessarily comparable. Raw scores cannot be interpreted easily because there is no meaningful zero point and no equal-interval scale. Standardized, norm-referenced tests, therefore, cannot produce accurate, easily interpretable measures of achievement in reading comprehension.

Sensitivity

Rather than setting out to assess gain within individuals, standardized norm-referenced tests are designed to "measure the stable, between-individual differences that traditionally have been of primary interest to psychological testing" (Carver, 1974, p. 512). The design principles of such tests, that is, deliberately maximize individual differences. For example, questions that most students answer either correctly or incorrectly are eliminated from the tests in the experimental stages. The most efficient question, for purposes of differentiating between individuals, has a passing proportion of .50 (or .625 when corrected for guessing). The tests, then, are referenced to a norm group rather than to an absolute criterion or a criterion based on specifiable test content; they are "so constructed that at each grade level they attain a normal distribution of test results" (Singer, 1973, p. 4). The reliability is determined by internal consistency and the stability of response to the same test administered at two different times. Any sensitivity norm-referenced tests might have for measuring gain or growth within individuals over a period of a school year is systematically eliminated in the item-selection process. Standardized, norm-referenced tests, that is, are insensitive to short-term achievement in reading comprehension. As a consequence, they are also insensitive to differences in educational treatments.

Format

Commercial firms design and develop most of the standardized, norm-referenced tests that schools depend on. The design, construction, and validation of these tests is time-consuming and requires considerable expertise, as well as what some commentators (e.g., Davis, 1964) call "artistry", so they are, of course, expensive testing instruments.

Part of the salability of these costly tests lies in their format: they come in pre-assembled packages that are easy to administer. But it is precisely that inflexible format which is the source of their limited utility and, as a consequence, their enormous hidden cost. The rigid format, for instance, containing only a few parallel test forms, permits only one simple evaluation design, a pre- and a post-test. Moreover, because the prepackaged tests cannot be taken apart and reassembled to construct a test of appropriate difficulty for an individual student or a particular group of students, standardized, norm-referenced tests yield imprecise measures of achievement. In order to measure student achievement in reading comprehension accurately, the test administrator must assign the student to a test form with a level of difficulty which is very close to the student's actual level of reading achievement. The more the test varies in difficulty from the student's actual reading ability, the more imprecise the measure of that ability. Since standardized, norm-referenced tests are inflexible in format, since they contain few parallel test forms, and since each form covers many levels of difficulty (e.g., a 4th grade student may face 10th grade reading materials), it is nearly impossible to measure an individual student's reading achievement accurately. Rigid test formats, then, are not only inherently expensive, but they prevent school systems from implementing satisfactory evaluation designs.

Summary

Standardized, norm-referenced tests of reading comprehension are reliable predictors of academic success, but they are entirely inadequate as measures of ability or achievement in fundamental comprehension skills. Though the publishers of standardized comprehension tests are loath to specify what such tests measure, factor analyses, high correlations with intelligence tests, passage independence of test items, and reviews of the content of the tests reveal a bias toward critical and evaluative reading skills. In other words, standardized comprehension tests slight what is usually called "literal comprehension"--those very abilities (1) that are basic to more advanced reading comprehension skills, (2) that take up a considerable portion of the reading and instructional time in most reading programs, (3) and that are vital to the development of a literate populace, a basic goal of school systems.

Indeed, if "levels" of comprehension (e.g., reading the lines, reading between the lines, and reading beyond the lines) are conceived as steadily expanding contexts for interpretation of the text or increasingly extensive relationships between the information in the text and the cognitive structures of the reader, it can be argued that there is little possibility of ever locating more advanced comprehension skills in that continuum until the base line is drawn, until literal comprehension is defined and tests of it thoroughly validated. Until then, tests of critical and evaluative reading skills (i.e., standard comprehension tests) are condemned to float indefinitely in the limbo of vague, global conceptualizations which are antithetical to the movement toward teaching by objectives and mastery learning. For, unless test-makers can identify the lowest level of meaningful synthesis (e.g., "literal comprehension") between the linguistic features of the text and the cognitive structures of the reader, what possibility is there for identifying more extensive and complex interrelationships?

In addition to these conceptual and theoretical difficulties, standardized comprehension tests have limited utility due to a lack of objectivity in test construction; scaling properties that make test scores difficult to interpret; insensitivity to gain within individuals and differences in educational treatments; and rigid, costly formats.

It is apparent that school districts need a test of literal comprehension based upon an explicit, viable conceptualization of literal comprehension. Further, such a test must be objective in construction, scaled with equal intervals and a meaningful zero point, sensitive to gain within individuals and differences in instructional treatments, and flexible in format.

CHAPTER II

COMPREHENSION[1]

Any attempt to measure reading comprehension should begin with a conceptualization of comprehension that is grounded in conventional usage.[2] People often use the phrase, "reading comprehension," to refer to the act or process of apprehending the meaning of written discourse. Since the process of comprehension is complex, extremely rapid, and entirely covert, the test-maker is necessarily limited to attempts to measure the product (rather than the process) of comprehension. And the product of comprehension, the thing to be apprehended, is meaning. As noted in the preceding chapter, a test of reading comprehension must measure a student's apprehension of the meaning of a particular segment of printed discourse. The obvious implication is that the test-maker must first identify the meaning (Carroll, 1972) or, more generally, the kinds of meaning (e.g., explicit) that are to be apprehended.

But the theoretical problems involved in identifying the meaning to be apprehended (much less measuring the student's apprehension of it) are labyrinthian. "Meaning" is even more conceptually ambiguous than "comprehension," and many a theory of comprehension, as Smith (1971) so wryly notes, has foundered on efforts to determine what "comprehension and meaning 'really are'" (p. 185). But the labyrinth seems to be unavoidable. Efforts to evade conceptual difficulties with "operational" definitions of comprehension have not resulted in viable tests of reading comprehension. Besides, what possible justification is there for labeling a test "reading comprehension" without "marshalling evidence in the form of theoretically relevant empirical relations to support the inferences that an observed response consistency has a particular meaning" (Messick, 1975, p. 955)?

---

[1] The following discussion of comprehension and meaning is based on a model of reading as a constructive language process, the most recent expression of which is Smith (1975). For a review of the evidence for such a model, see Ryan and Semmel (1969). Katz (1972) was the primary source for the competence model assumed by the performance model.

[2] "Ordinary language often embodies concepts which have developed and endured because they capture something of significance to human beings. Thus, ordinary language concepts have, at least, a prima facie right to our consideration, especially when we are studying human beings, and they should be replaced by a technical vocabulary only when there are clear empirical advantages in doing so and when we are clear about the human significance of the change introduced" (Strike, 1975, p. 462).

Operational Definitions

The point is important enough to warrant an extended example. Attempts to avoid pursuing the psycholinguistic ramifications of a given test of comprehension often result in "operational" definitions that defy conventional usage and consequently promote misunderstanding in a field already rife with ambiguous concepts. In the study by Bormuth, Manning, Carr, and Pearson (1970), for example, "a comprehension skill is defined as the ability to respond correctly to a question beginning with the letters 'wh' which deletes one of the immediate constituents of a syntactic structure" (p. 351). Now obviously teachers traditionally ask such who-what-which-where-when-how-why questions in order to direct attention to important features of the text under scrutiny and to promote "comprehension skills," but when the text is available for perusal as it is on a comprehension test, a student with minimal syntactic competence can locate the correct answer to such questions without necessarily understanding what the sentence means (Anderson, 1972; Carroll, 1972). The limitations of the wh-item as an operational definition of comprehension are evident in the following nonsense sentence:

The izilbe gatgotted the dizzleboo. Who *gatgotted the dizzleboo*? Obviously the *izilbe*. But what does the sentence mean? And the problem is not exaggerated by nonsense sentences. Consider the following statement: *Incantatory glee reverberated in Johann's miniscule cerebrum*. Students with a minimal syntactic competence and a little test-wiseness could locate *incantatory* in the text without the vaguest idea of what *incantatory* or the rest of the sentence means. As students become familiar with such test items, they should be able to locate the right answer in the text long after the reading passages exceed their vocabulary knowledge. Verbatim transformations of sentences in a text, therefore, have a limited life-span as viable tests of comprehension.

Paraphase questions, on the other hand, rather than solving the difficulties inherent in such "transformed verbatim questions," only reintroduce some of the same theoretical problems that plague standardized comprehension test items. "Any change in wording, including substitution of synonyms, usually alters [the] meaning" of the original text (Johnson, 1975, p. 429; Alston, 1964; Lyons, 1968; Quine, 1960; Smith, 1975, p. 104); vocabulary changes, that is, introduce the test-maker's own idiosyncratic interpretation of the text--his *approximation* of what the text "means"--into the test items. *Mirth*, for example, simply does not mean the same thing as *glee* in the "*incantatory*" sentence above. Even simple active and passive transformations engender different understandings (Johnson, 1975, p. 437; Anisfeld and Klenbort, 1973; Herriot, 1970; Offir, 1973; Smith, 1975, p. 104). (In light of the apparent inability to change the wording of the text without "engendering different understandings," the very concept of a paraphase--changing the words while maintaining the same meaning--seems self-contradictory.")[3]

---

[3] The notion of meaning here is extended to cover "construal" and "stylistic" features of an utterance (cf. Katz, 1972).

In addition, the inflated syntax of some of the compound wh-items described by Bormuth et al. (1970) tend to make the questions more difficult to understand than the original sentence in the text, a common problem with standardized tests. For example, from the sentence, He (the boy) fractured his arm, the questions, Who was it who fractured his arm? and Who was it who broke his arm? are derived (p. 352). (It is also worth noting that, in Anderson's [1972] opinion, only correct responses to paraphase questions among wh-items can be adduced as evidence of comprehension, yet the examples above, even though they are labeled "paraphase," fail to conform to Anderson's definition: Two statements are paraphases of each other if "1) They have no substantive words [nouns, verbs, modifiers] in common and 2) they are equivalent in meaning" [p. 150]. In the paraphase transformation quoted above, however, only the verb is changed--*fractured* is replaced by *broke*.)

In summary, correct answers to verbatim transformations cannot be cited as sufficient evidence for comprehension because it is possible to answer such questions correctly without comprehending the sentences upon which they are based. The operational definition of comprehension, that is, does not "preclude plausible rival interpretations" (Messick, 1975, p. 959). Paraphrase transformations, on the other hand, are subject to many of the same criticisms that are leveled at standardized comprehension test items.

The problem is how to write test items that are impossible to answer (beyond guessing) without apprehending the meaning of the text upon which the questions are based. The brief critiques of the wh-items and standardized comprehension tests in this report should make it evident that test-makers are caught on the horns of a dilemma: If they avoid imposing idiosyncratic meanings on the text by writing test items based on minimal transformations of the text, then it is possible to answer the questions without apprehending the meaning of the text. On the other hand, if test-makers change the wording of the text in any way in the test items in order to force the student to interpret the text, then they impose idiosyncratic interpretations on the text without randomly sampling all possible interpretations of it and introduce unnecessary difficulties in the syntax and vocabulary of the test items. How then is the meaning which students are to apprehend to be identified without prejudicing it? Further, is it possible to conceptualize meaning without getting bogged down in the "interminable controversies...about what kind of thing meaning is" (Katz, 1972, p. 1)?

Meaning

In light of the criticism that both standardized comprehension tests and wh-items (paraphrase transformations) impose idiosyncratic interpretations on the text, it appears to be crucial for test-makers to identify, insofar as possible, the relationship between meaning and the orthography on the printed page rather than to speculate on the absolute nature of meaning, since such speculations inevitably collapse into philosophical quibbles. For the limited purposes of this discussion, the relationships between meaning and the text are reduced to three simplified possibilities: (1) Meaning is derived from the text; (2) meaning is imposed upon the text; or (3) some combination of (1) and (2).

Deriving meaning from the text.

The first possibility--that meaning is derived from the text--implies that meaning is in the text, or, more exactly, that meaning is in "language" and represented rather accurately by the orthographic system on the printed page.[4] Thus, a transformational grammarian might contend that the meaning of a discourse is a result of the "grammatical and semantic relations which obtain within and among the sentences of the discourse" (Katz and Fodor, 1967, p. 172). In Katz's (1972) semantic theory,

---

[4] Phonological rules may be bypassed in the interpretation of written discourse (Venezky, 1967). Chomsky and Halle (1968) also point out that meaning is more directly represented in the orthography on the page than it is in the phonological component of language (e.g., sane, sanity). "There is an essentially arbitrary relationship between sound and meaning so that properties of phonetic shape do not predict properties of propositional form and vice versa" (Katz, 1972, p. 367).

> the semantic component of a grammar contain[s] a dictionary[5] that formally specifies the senses of every syntactically atomic constituent in the language. It [i.e., the semantic component] must also prescribe rules for obtaining representations of the senses of syntactically complex constituents, which are formed from representations of the senses of their atomic constituents in the dictionary. The dictionary provides the finite basis and the rules provide the machinery for projection onto the infinite range [of the possible combinations of the senses of the lexical items]. (Katz, 1972, p. 33) The idea underlying this conception is that the logical form of a sentence is identical with its meaning as determined compositionally from the senses of its lexical items and the grammatical relations between its syntactic constituents. (p. xxiv)

In this "compositional" account of meaning, the semantic component of the grammar

> operate[s] exclusively on the underlying phrase markers in the description of a sentence.... Semantic interpretation proceeds, first, by an assignment of lexical readings from the dictionary to the atomic constituents of a sentence and, then, by an assignment of derived readings to each syntactically complex constituent by the operation of the projection rule upon the readings of its component parts. (Katz, 1972, p. 415)

Thus, initial syntactic analysis--identification of underlying phrase markers--is prior to (in the sense of directionality) the interpretation of the deep structure (underlying phrase markers) of a sentence: "The syntactic component is the generative source of a grammar. Its output is the input to both the phonological component and the semantic component" (p. 31).[6]

---

[5] This is of course the "ideal" dictionary, not to be confused with that tribe of paper dictionaries exemplified by the Oxford English Dictionary.

[6] There is some debate among transformational grammarians over the interpretation of final derived phrase markers by the semantic component (Chomsky, 1970).

Now this is an attractive theory for a test-maker because it seems to allow for the derivation of meaning from a given text by a finite set of meanings combined by a rule system which can be explicitly stated, thus raising the possibility of objectively deriving and specifying all possible interpretations of a text. Meaning is therefore in language, free from the dispositional limitations of any reader who might encounter language in one of its empirical manifestations. The medium itself is never at fault in any failure to encode or decode meaning accurately: "Each human thought is expressible by some sentence of any natural language," and failures to express or derive meanings accurately are not attributable to failures in the expressive capacities of language but rather to an individual's lack of skill "in exploiting the richness of his language" (p. 19).

But this brief outline of Katz's semantic theory should make it evident that such a possibility for objectively deriving meaning from a discourse is based upon a competence rather than a performance model. Katz's model is erected on the notion of sentence types rather than tokens:

> We based our study of the meaning of sentence types on an idealization that allowed us to focus exclusively on linguistic meaning by abstracting away every aspect of language that does not reflect pure grammatical competence. We observed early in the book that even a complete theory of the meaning of sentences and other constituent types is a far cry from a full theory of linguistic communication. (p. 443)

The test-maker, however, cannot ignore the communicative limitations of the reader since they affect the response consistency of the test and are, therefore, precisely the point of interest of the test-maker (and the teacher).

Katz distinguishes between a competence and a performance model as follows:

> In the theory of linguistic competence we seek to state the system of rules that formally represents the ideal linguistic structures that underlie the utterances of natural speech. We idealize away from the distortions and irregularities characteristic of natural speech and concern ourselves with the systemization of those aspects of natural speech that directly reflect the contribution

> of a speaker's fluency. The theory of linguistic performance, on the other hand, seeks to account for the principles that speakers use in actually producing and understanding natural speech. Accordingly, the study of performance assumes the contribution of competence and directs its attention to the manner in which the contributions of various psychological factors--e.g., memory limitations, attention shifts, distractions, brain damage, errors--interplay with linguistic factors to produce natural speech, with all its characteristic distortions and irregularities. (p. 25)

Though a performance model "assumes the contribution of competence," testmakers cannot wait for the definitive competence model (which Katz projects into the next century). What is needed is a "working" model of the manner in which readers apprehend the meaning of connected discourse, taking into account the dispositional limitations of the reader and the differing interpretations of a given text resulting from the various verbal and extraverbal contexts in which it occurs. The first consequence of shifting from a competence to a performance model, however, is to lose the ability to specify the meanings to be apprehended.

Imposing meaning upon the text. Psychologists, in marked contrast to transformational grammarians, usually maintain that meaning is in the reader rather than in the text or "language." Osgood (1967), for example, argues that

> The meaning which individuals have for the same signs will vary with their behaviors toward the objects represented. This is because the composition of the mediation process, which is the meaning of a sign, is entirely dependent upon the composition of the total behavior occurring while the sign-process is being established. (p. 163)

Thus, in developing a model of reading as a constructive language process, Smith (1975) locates meaning not in "language" but in "the underlying thought processes of the language user" (p. 84). According to Smith, it is impossible to derive meaning from a text because "there is no one-to-one correspondence between the surface and deep structures of language" (p. 84).[7] Meaning is first imposed upon language in the deep structure prior to syntactic analysis or, for that matter, prior to sampling any of the linguistic clues to the meaning intended by the writer. That is, a reader makes an hypothesis about what any given sentence in a discourse means based upon his expectations which are created by the general sociolinguistic situation in which the discourse occurs, the meaning imposed upon the preceding sentences of the discourse, etc. Having made his initial hypothesis, the reader than samples selectively amongst the linguistic clues to meaning in the text. If the original hypothesis is verified by the information he perceives in the text, the reader moves on to the next sentence. If the original hypothesis is not substantiated by the information in the text, then the reader either samples more extensively or changes his hypothesis about what the sentence means and samples again.

Now such a notion of meaning that is initially separate and distinct from the linguistic clues to meaning in language certainly confronts the full dispositional limitations of the reader and the various contextual features in which the discourse occurs, but it is impossible for the test-maker to identify "the meaning" to be apprehended, for meaning is essentially and ultimately idiosyncratic. That is, "comprehension," in Smith's performance model, refers to the assimilation of the information[8] in the text to the cognitive structures of the reader. Given the location of meaning in the cognitive structures of individual readers, it follows that "the meaning"

---

[7] "One reason that the surface structure of language does not have a one-to-one relation with the underlying deep structures of thought is that case relations can be represented in a variety of ways" (Smith, 1975, p. 103).

[8] Perception of parts of the orthography of the text as "information" (rather than "noise") is itself an act of comprehension (Smith, 1975).

of a particular segment of printed discourse varies as a function of the disparity between the cognitive structures brought to bear on the discourse. Different readers, as any student of literature knows, interpret the same text in different ways.[9] Alternate possibilities for interpreting an utterance often surprise a reader/listener, which is only an indication that the decoder's perceptions about the utterance are restricted by his own cognitive "set." Moreover, any reader comes to the same text on different occasions with varying moods, degrees of attentiveness, purposes, presuppositions, available knowledge, etc., all those personal idiosyncracies eschewed by a competence model (Katz, 1972, p. 15). That is, the "array" of cognitive categories that any reader can bring to bear on the information in the text varies with the dispositional limitations of the reader. Therefore, the interaction between the information in the text and the cognitive structure of the reader varies not only between readers but also within readers.

Not only is meaning (theoretically) idiosyncratic, but it may also be non-verbal and non-observable. As noted previously, Smith (1975) contends that meaning lies in the thought processes of the language user and that there is no one-to-one correspondence between meaning and the surface structure of language. Pursuing the notion further, Smith (1971) is forced to characterize "the meaning of a sentence [as] something global, a 'state of mind,' an instantaneous set of relationships established in the cognitive organization" (p. 194). Meaning is merely the absence of uncertainty (Smith, 1971, 1975).

Now conceptualizing meaning in terms of the non-verbal, non-observable dispositional idiosyncrasies of the reader does not in itself preclude measurement. Psychologists are long used to measuring dispositional phenomena that are non-verbal and non-observable (and that sometimes do not exist except in the imaginations of psychologists). Brown (1958), for example, writes that "a disposition is discovered by creating various contingencies and observing responses" (p.103). But such measurement techniques are rudimentary and have never proven very successful, even in dealing with single words, much less the complex interrelationships among the words of a sentence (Miller, 1965).

---

[9] The "definitive" reading of a text is a parochial notion, always deflated in time.

By pursuing the full implications of the performance model of reading as a constructive language process, the test-maker is left in a considerable quandary: How can the meaning(s) of a segment of connected discourse to be apprehended by the student be identified if they are infinitely variable, non-verbal, and beyond the capacity of psychometricians to measure? Further, if meaning is non-verbal and there is no one-to-one correspondence between meaning and surface structure, then what appears on the printed page is never more than an approximation of the meaning as intended by the writer or the meaning apprehended by the reader. The speaker or writer straining to say or write what he "really" means comes immediately to mind. Katz's ascription of the failure to express a thought accurately to the user and not to language is turned around here; since meaning is not in language, and there can be no efficient transfer of meaning from the simultaneity of non-verbal cognitive structure to the temporal realization of meaning in a string of morphemes, the failure to express a thought accurately lies finally in the medium rather than in the language user.

If meaning is essentially non-verbal, and idiosyncratic representations of meaning in verbal form are never more than approximations of "the meaning" intended by writers or "the meaning" as apprehended by readers, then it is impossible for a test-maker to identify the meanings to be apprehended by the

the student. Efforts to list all possible interpretations of a text and then to sample randomly from that list are fundamentally misconceived. Given that it is impossible to specify all the acceptable meanings of a text, is it possible for a test-maker to identify, in general, the "kinds" of meanings (e.g., explicit) to be apprehended? Further, is it possible to combine some of the features of Katz's competence model, which allows for the identification of both specific meanings and types of meaning, with features from Smith's performance model, which allows the test-maker to identify the dispositional limitations of the reader and the context within which a text is interpreted? Finally, is it possible to specify the "level" or "degree" of comprehension (e.g., literal) indicated by a particular response type?

Explicit meaning. Teachers often identify meanings as explicit or implicit, literal or inferential, etc. If such distinctions are viable, then it is possible to specify the kinds of meanings to be apprehended at a given "level" of comprehension. For example, literal comprehension can be defined--that is, located in relation to other "levels" of comprehension on one side and in relation to non-comprehension, perhaps "mere verbalization" or "recognition," on the other--as the apprehension of the explicit meaning(s) of connected discourse. The preceding discussion should make it evident, however, that there is no "explicit" meaning in the text even though people speak (metaphorically) of what the text "explicitly says." Clearly the text does not "say" anything; all meaning is implied or inferred or derived from or imposed upon the linguistic clues to meaning in the text.

The explicit/implicit dichotomy in meaning seems to be founded upon the distinction between denotative and connotative meanings. According to Webster's New Collegiate Dictionary (1974), denotation refers to the "direct, specific meaning" of a word or what is commonly called its referential aspect.

The preceding discussion, however, casts considerable doubt on the notion that words, much less sentences, have any "direct, specific meanings" that can be represented in the surface structure of language. "Denotation" and its companion concept, "explicit meaning," are rooted in a failure to distinguish between reference, usually attributed to words in isolation, and meaning, which always accrues to words in complex interactions with other verbal and non-verbal experiences. Even if the referent of a word is identified as a "psychological entity" (Johnson, 1975, p. 426), thus blurring the distinction between denotation and connotation, there is still no one-to-one correspondence between the referential associations in the brain/mind of the reader and the orthography on the printed page. The clues to meaning in orthography are minimal[10]--simple temporal sequences representing the complex simultaneity of cognitive structures. Moreover, "denotation" seems to result from a habit of analyzing words in isolation (as if words ever existed in "isolation") and leads to the false assumption that the meaning of a sentence is the sum of the meaning of its parts (Miller, 1965). "A speaker's ability to understand any sentence depends in part on his knowing the meanings of its component morphemes" (Katz, 1972, p. 35), but a "morpheme" is quite a different notion from a "word," which may only be an artifact of the orthographic system (Smith, 1975). Besides, "the same set of morphemes can mean different things when put in different syntactic arrangements" (Katz, 1972, p. 35): e.g., Philbert is munching on a crawdad / A crawdad is munching on Philbert.

Critics, psychologists, and linguists have long inveighed against treating words as entities whose meaning could be isolated from the dynamics of the contexts in which they occur. I. A. Richards (1936/1965), in what amounts to a precursor of the "current" model of reading as a constructive language process or a "psycholinguistic guessing game" (Goodman, 1970), criticizes the attempt to take

---

[10] As noted previously, however, meaning may be represented more clearly in orthography than phonetics.

> the senses of an author's words to be things we know before we read him, fixed factors with which he has to build up the meaning of his sentences as a mosaic is put together of discrete independent tesserae. Instead, they are resultants which we arrive at only through the interplay of the interpretative possibilities of the whole utterance. In brief, we have to guess them and we guess much better when we realize we are guessing, and watch out for indications, than when we think we know. (p. 55)

Brown (1958) also contends that

> an attempt to understand the meaning of a single linguistic form in isolation from the total language process would be rather like trying to understand a single bid in isolation from a game of bridge. The meaning of a form, its total conventional usage, involves the full language game. (p. 106)

Chafe (1972) pushes the interrelatedness of the component parts of speech even further: "The point is that we do not use only part of what we know when we say something, we use all of it, and there is no way to divide knowledge that is linguistically relevant from knowledge that is not" (p. 67). An analysis of the particular senses of the meaningful units of discourse mushrooms quickly into a theory of knowledge.

Holistic meaning. It has been the contention of students of language ever since Aristotle that meaning is holistic and that the sentence carries the primary burden of meaning in discourse. Teachers, for instance, make distinctions between "reading the line, reading between the lines, and reading beyond the lines," which seems to be a more viable categorical scheme than the denotative/connotative dichotomy simply because it deals with whole sentences rather than words in isolation. Reading the line, reading between the lines, and reading beyond the lines suggest that there is an expanding context--intrasentential, intersentential, and extrasentential--within which the information on the printed page can be interpreted or, from the point of view of Smith's performance model, that there is an increasingly extensive set of cognitive categories to which the information on the printed page can be assimilated.[11] If the sentence is identified as the primary vehicle for conveying meaning in written discourse, there seems to be some possibility of identifying the kinds of meanings to be apprehended, that is, the identity and extensiveness of contextual constraint on the clues to meaning in and beyond the text and the identity and extensiveness of the cognitive structures to which that holistic information unit in the text has to be assimilated. Thus the key to a synthesis of the specificity and objectivity of Katz's competence model with the ability to account for the dispositional limitations of the reader made possible by Smith's performance model is Katz's assertion that "the empirical existence of a natural language lies in the linguistic rules internalized by its speakers" (p. 15) and Smith's (1975) notion that "language is [always] embedded in meaning" (p. 105).

For it is obvious that, in spite of the idiosyncratic, non-verbal nature of meaning, speaker/writers and listener/readers do in general come to some agreement about the meaning(s) that each of them apprehends in a given message as indicated by their response behavior to the message. (Gross misunderstandings are usually due to egregious errors in encoding the message--i.e., misapplications of the shared psycholinguistic rules system--or a misapprehension of the context within which the message is embedded--i.e., a misapplication of the shared sociolinguistic rules system.) Though surface structures may only be approximations of the deep structures of language or the "abyssal" structures from which meanings may be generated, a well written text clearly allows for some general agreement about what the text means else books would not have become as pervasive as they have in their brief [illegible]-

---

[11] This is parallel with Johnson's (1975) notion that the "meaningfulness" of a word is determined by "the extensiveness of the network of referential associations" (p. 427).

association with language.

What is so remarkable about language comprehension is that people do understand each other, that the apprehension of the multiplicity of meaning inherent in any relationship between utterance and decoder is in practice such a rare event that it is more often a source of amusement (Smith, 1975, p. 105) than dismay. Indeed, those people who develop skills in mining the inherent multiplicity of meaning in surface structure are more often considered verbal "artists" than malaprops.

Commonality of meaning. What is the source of the apparent commonality of meaning that can be apprehended in well written texts? It is interesting to note that disputes about what a given text "means" are usually referred back to what the text (metaphorically) "says" or, more specifically, to the orthographic features on the printed page. Katz (1972) attributes this commonality of meaning derived from a text to the regularity of language:

> If the way in which the speaker finds the words with which to express his thoughts is not, at least in part, the same way that his hearer recovers the thought from the articulated words, the fact that different speakers of the same language can freely exchange positions as speaker and hearer, always associating the same thought [?] with the same sentence, would be incomprehensible. Therefore, the basic question to ask is what are the common principles for encoding and decoding. (p. 24)

Smith (1975), on the other hand, following the generative semanticists, goes beyond language to the contingent circumstances in which an utterance occurs to account for commonality of meaning:

> The meaning of an utterance involves much more than the words spoken; it depends on the entire situation, verbal and non-verbal, in which the utterance is made.... Language is embedded in meaning, and meaning is always limited by the prior purpose and understandings of both speaker and listener, or writer and reader. (p. 105)

Sociolinguists, for example, have contributed greatly to understanding how little linguistic information it takes to convey complex meanings in carefully defined social situations.[12] In a like manner, Freedle and Carroll (1972) also contend that

> Understanding language nearly always involves not only comprehending the words and grammatical structures of a message as linguistic symbols, but also taking account of those knowledges, facts, or ideas that underlie the message but are not explicity built into it....Much of the semantic content of discourse is not to be found in the spoken or printed words themselves, but in the prior knowledge that the producer of a message assumes the hearer or reader to have. (p. 360)

These attempts to account for common interpretations of the same text reflect (at least) three separate notions of the relationship between meaning and sentences in the text. (1) Compositional meaning: The meaning of a sentence (type) is determined by the meaning of its constituent parts and the grammatical interrelationships among them. Such a notion accounts for synonymy, paraphrase, etc., but is insensitive to context and the dispositional limitations of the reader. (2) Contextual meaning: The meaning of a sentence is determined by the interrelationships among the compositional meaning(s) of a sentence type and the context in which it occurs as a token. ("The upper limit of semantic interpretation in a grammar concerned with conventional or linguistic meaning [i.e., compositional meaning] is the starting point for a theory of contextual construal" [Katz, 1972, p. 445].) Such a notion still accounts for synonymy, paraphrase, etc., assuming that context can be specified, but is insensitive to the dispositional limitations of the reader. (3) Dispositional meaning: The meaning of a sentence token is determined by the interrelationships among the compositional meaning(s) of the sentence type, the specific context in which the sentence type occurs as a token, and the dispositional limitations of the reader. "Dispositional meaning" amounts to an internalization of both compositional and contextual meaning. There is no other way to account for the dispositional limitations of the reader in a performance model. In a performance model, meaning is a function of the interaction between the features of the text and the context as perceived by the reader. Meaning is in the reader, and what can be observed in verbal or non-verbal response to the text is only an indication of the meaning apprehended by the reader. Such a notion will account for some commonality of meaning apprehended by readers as indicated by response consistency to the text, but will not account for the kind of specificity of meaning implied by "synonymy," "paraphrase", etc. since an extra-linguistic account of meaning which is peculiar to the reader is interacting with compositional and contextual meaning. Commonality of meaning is ultimately attributable to similarities in cognitive structures among readers.

---

[12] See Bernstein (1969).

Note that compositional meaning is integral to all three accounts of meaning above, but neither contextual nor dispositional meaning is integral to compositional meaning (unless the latter is considered an expression of the dispositional capacities of the reader). Note further that those aspects of language which may be genetically coded--e.g., a tendency among natural languages toward similar syntactic structures (Chomsky, 1968; Lenneberg, 1967)--lie also within the compositional account of meaning. It is tempting to explain all commonality of meaning as compositional; indeed, Katz does so, using such terms as "literal," "linguistic," "conventional," and "compositional," interchangeably. Hence, literal comprehension could be defined as the apprehension of the compositional (i.e., literal, linguistic, or conventional) meaning of the discourse, and compositional meaning could be identified quite accurately as "the grammatical and semantic relations which obtain within and among the sentences of the discourse" (Katz and Fodor, 1967, p. 172). Those other "levels" of comprehension--reading between and beyond the lines--which always lead to increasing diversity of interpretation could then be distinguished quite precisely from literal comprehension.

But, as noted previously, any conceptualization of comprehension that does not account for the dispositional limitations of the reader has a limited utility for test-makers (and teachers). Comprehension is the apprehension of meaning, and "apprehension" demands an account of the dispositional limitations of the reader. Internalizing syntactic and semantic competencies in the cognitive structures of the reader does not solve the problem either. No act of apprehension of the meaning(s) of a sentence ever occurs free of contextual contingencies. Meaning is always embedded in meaning (Smith, 1975, p. 105). Any "level" of comprehension, therefore, involves compositional and contextual meaning in dynamic interplay with the dispositional limitations of the reader. The testing situation offers a unique opportunity to identify and control the interactions between those three aspects of meaning.

Measuring Comprehension

Any attempt to measure a student's apprehension of the meaning of written discourse introduces two additional factors--the item type and the testing situation--into an already complex cognitive process. Measurement of a process nearly always disrupts the process to some extent, and the process reflected by the measurement procedure is partly peculiar to that procedure. This is certainly true of the measurement of comprehension. The meaning apprehended by a student on a reading comprehension test is a function of the interaction between the text, the item type, the testing situation, and the student. The failure to identify and control interacting features of the test inevitably results in rival interpretations of response consistencies to the test. It was argued in preceding sections of this proposal, for instance, that correct responses to items on standardized comprehension tests were not evidence of comprehension of the passages in question because the test items were not passage dependent, i.e., the interactions between item type and text were not defined and controlled. It was also argued that correct responses to verbatim transformations of sentences in the text were no evidence for comprehension because the carefully controlled interaction between text and item type excluded meaningful aspects of the discourse; i.e., the processing of text was primarily syntactic rather than semantic. Finally, it was argued that paraphrase transformations of sentences in the text and test items on standardized comprehension tests introduced the test-writer's own idiosyncratic interpretation of the text into the test items and often made the test items more difficult to comprehend than the text itself. Again, the problem was a failure to define and control the interaction between text and test item.

A correct response to a particular item type can be accepted as evidence of comprehension of the text upon which the item is based only if it can be demonstrated that the correct response is impossible (beyond chance) without apprehending "the grammatical and semantic relationships which obtain within and among the sentences" of the text. Passage dependency, in other words, is the first demand to make of any item type. If the item type is not passage dependent, then there is no further possibility of defining the interaction between test item and text. Indeed, there may be none. The test item must bear a specifiable relationship to both the syntactic and semantic features of the text; in addition, the extensiveness of that interaction--e.g., intrasentential, intersentential, and extrasentential--must be identified and controlled before the test can be labelled as to the "level" or "degree" of comprehension it assesses (e.g., "literal" comprehension). This latter constraint on test construction amounts to a specification of the context within which the information in the text is to be interpreted (e.g., does the item type demand information other than "the grammatical and semantic relations that exist within and among the sentences of the discourse," and, if so, where does this information come from, who is expected to have access to it, and what skills and processes are involved in integrating that extra-textual information with the text?).

Moreover, any interaction between student, test situation, item type, and text involves assumptions about requisite competencies on the part of the student which must be matched properly by the test tasks, otherwise response consistencies are again difficult to interpret. Texts vary greatly in syntactic complexity, for example. How is the syntactic complexity of the text to be ascertained and controlled in relation to the syntactic abilities of the students taking the test? Is a student to be declared incapable of relating inter-textual and extra-textual information meaningfully when the text itself already exceeds his ability to apprehend the grammatical relations that exist within and among the sentences of the text? What level of linguistic competence is assumed by the test? How are general linguistic and intellectual abilities to be differentiated from those abilities that are peculiar to the item type?

Assuming that the student has the requisite competencies to perform properly on the test, how is the test to be administered so as to eliminate, insofar as possible, the non-requisite competencies (e.g., phonetic skills) from the test scores? How can the test be designed and administered to reduce the effect of personality, motivational factors, test-taking skills, etc.?

Since traditional item types (i.e., questions based upon the text) make the interaction between item type and text so difficult to identify and control, the obvious solution to that problem is to eliminate questions. The following two chapters analyze the cloze procedure as a test of comprehension without questions. An attempt is made to specify the interactions between text and item types on several variations of the cloze procedure. Chapters on validity later in this report attempt further specifications of the interaction between text, item type, testing situation, and student characteristics.

CHAPTER III

THE CLOZE PROCEDURE

Wilson Taylor introduced the cloze procedure to the reading field in 1953 as "a new tool for measuring readability." Taylor derived the term "cloze" from the concept "closure" in Gestalt psychology, reasoning that "the human tendency to complete a familiar but not-quite-finished pattern" is comparable to supplying missing words in connected discourse (p. 415) Though Taylor's analogy with Gestalt concepts was misleading (Rankin, 1964; Weaver, 1965; Ohnmacht, Weaver, and Kohler, 1970), most of his procedures and conclusions about the cloze procedure have proven remarkably durable through more than 20 years of cloze research. In addition, the "new tool" that Taylor introduced to measure readability has been extended dramatically in investigations of "reading comprehension, learning, information, thinking, numerous language variables, teaching, aptitude, readiness, listening, flexibility, and context cues" (Rankin, 1974, p. 2).

A complete bibliography of cloze research would be comprised of several hundred items. What follows is a brief, critical review of selected studies, concentrating on salient features of the cloze and related theoretical issues which are germane to the analysis of comprehension as discussed in the preceding chapter of this report. For more comprehensive reviews of the literature on the cloze, the reader is referred to Rankin (1959, 1965, and 1974), Potter (1968), and Fram (1972).

Readability

Readability formulas.[1] Conventional readability formulas measure a small number of variables such as prepositional phrases per 100 words, percent "hard" words, and average sentence length in specified segments of a text and then calculate scores which indicate the grade level or levels at which students with average reading abilities will be able to comprehend the text. The formulas were derived by analyzing written texts for which grade levels had been established on the basis of pupil performance and then determining by regression analysis the relative weightings of sentence length, hard words, and so forth that would best "predict" the grade level or difficulty of the texts.

Once established, the formulas were used to predict the grade level of other texts. They give teachers and publishers an estimate of the readability or difficulty of written material without actually having students read it. However, they have shortcomings in that they do not tell how individual students or groups of students will respond to specific texts, and they do not take full account of complexities of form and content which may affect the comprehension of individuals.

With few exceptions (notably Bormuth, 1966), readability formulas sample only two or at best three of the many stylistic variables that affect readability. The Lorge (1939), Flesch (1948), Dale-Chall (1948), and Spache (1953, 1960) formulas, for example, all count the average number of words per sentence but ignore variations in sentence structure, which can radically affect comprehensibility. For instance, scrambling the words in a sentence would not even affect the score on most readability formulas.

---

[1] See Klare (1974) for a current review of readability formulas.

The Lorge, Dale-Chall, and Spache formulas also count the "hard" words in a passage but are insensitive to the difficulty of words in context. Carroll (1971) has demonstrated that the comprehensibility of words can vary greatly with their grammatical functions (e.g., compare rank as a noun or verb to rank as an adjective). Moreover, stylistic elements vary in difficulty for students at different stages of language development, almost necessitating special formulas for each level of reading ability (Smith and Dechant, 1961). "Until the advent of the cloze test there was no practical way to measure the comprehension difficulties of individual words and sentences" (Bormuth, 1966, p. 85).

The cloze procedure. A standard cloze test of the readability of printed discourse is constructed in six easy steps (Taylor, 1953): (1) Delete every *n*th word (usually every fifth or more words)[2] irrespective of part of speech or meaning; (2) replace every missing word with a blank of standard size; (3) assign the "mutilated" passage to a representative sample of the students in question; (4) ask the students to fill in the missing words by guessing, from the remaining context, what the missing words might have been; (5) total the exact-word replacements[3] and calculate a readability score--the percentage of correct responses--on the basis of the total number of deletions; (6) compare the students' scores from different passages and rank the passages in order of difficulty.

---

[2] If the content surrounding a missing word is reduced below six to ten words, it becomes very difficult to replace the missing word (Aborn, Rubenstein, and Sterling, 1959; MacGinitie, 1961).

[3] Minor misspellings are accepted. Scoring synonyms, on the other hand, has little effect on test reliability or validity; instead, it introduces subjectivity, difficulty, and expense into the cloze procedure (Taylor, 1956; Bormuth, 1967a).

An estimate of the difficulty of every word or sentence in a passage can be obtained by constructing five forms of the test, deleting every fifth word, beginning alternately with the first, second, or third word, and so on, until every word in the passage has been deleted in one or another test form (Taylor, 1956; Bormuth, 1964). Different forms of the test are then randomly assigned to representative samples of the students and analysis made of performance on different forms.

Besides the ability to estimate the difficulty of every word and every sentence in a passage, the cloze procedure has several other advantages over readability formulas. First of all, a cloze test actually "measures" rather than predicts the readability of a passage. More specifically, the cloze procedure counts the number of successful, exact-word replacements of missing words in a passage and then expresses this number as a percentage of the total missing words. The percentage of correct responses indicates "the extent of likeness between the language patterns used by the writer to express what he meant and those possibly different patterns which represent readers' guesses at what they think the writer meant" (Taylor, 1953, p. 417). Thus cloze scores represent "the proportion of predictable material that the passage contains" (Coleman and Miller, 1968, p. 371) for the students in question. A student's ability to guess a significant proportion of the language used in a particular text indicates a sufficient acquaintance with the stylistic variables and the content of the text to be able to comprehend it with some specifiable degree of proficiency. Any teacher or subject coordinator, using the cloze procedure, can determine the appropriateness of a given text for a particular group of students.

Secondly, cloze scores reflect many more linguistic variables, including syntactic complexity (Ruddell, 1964; Simons, 1970; Stedman III, 1971) and various stylistic devices (Bormuth and MacDonald, 1965), than readability formulas. Taylor (1953), for example, compared the difficulty ratings of three passages as estimated by the cloze procedure and the Flesch and Dale-Chall readability formulas. The passages were ranked in the same order of difficulty by all three methods, but the cloze scores showed far more sensitivity to stylistic variables. Whereas the Flesch and Dale-Chall formulas predicted that a passage by Gertrude Stein would be appropriate for fourth or fifth grade students, the cloze procedure gave it a higher rating more consistent with its obvious difficulty.

In studying the validity of the cloze as an estimate of readability, Bormuth (1962) compared cloze scores on nine passages with multiple-choice and sentence-completion comprehension scores on the same passages. The correlation was .92.

Coleman and Miller (1968) used a modified cloze procedure to calibrate 36 passages for difficulty. Assuming that the amount of "new information" that can be gained from a passage is a function of its difficulty or the amount of predictable verbal material in a passage, they asked students to guess each successive word in the passages. If the student guessed the wrong word, he was corrected. Each student went through each passage twice, and the "information gained" was the difference between the two scores. Thus the final score reflected the difficulty of a passage or "the efficiency with which a passage transmits new information" (p. 369).

Aquino (1969), using the same 36 passages, compared Coleman and Miller's difficulty ratings to results from two other measures of "readability"--word-for-word recall and judgements of difficulty. The 36 passages were ranked in the same order of difficulty by all three methods.

The cloze procedure, however, can be a bit more cumbersome than readability formulas. For example, Dale-Chall readability scores are usually calculated with passages of at least 100 words in length. Using the cloze procedure, however, a reading teacher with 30 students would need a passage of at least 750 words to get a reliable estimate of difficulty. The reliabilities of cloze estimates of readability vary as a function of the number of students and the number of deletions (Bormuth, 1965). "Where the passage is very short (containing fewer than 30...[deletions]), it is doubtful that individual scores are sufficiently reliable to permit an accurate judgement of how well a given individual understood the passage" (Bormuth, 1967a, p. 16). Increasing deletions, and, consequently, passage length, tends to reduce error more effectively than increasing the number of students. As few as 40 deletions, or a passage of 200 words in length, however, could be used with 150 students (Bormuth, 1965).

Merely ranking passages for difficulty in relation to each other does not provide teachers with sufficient information about readability. In 1971, Bormuth attempted to develop "standards of readability" so that any cloze score on any given passage could be interpreted independently of other passages. He compared cloze scores to measures of "information gain" (the difference between pre- and post-test scores) assessed by multiple-choice and sentence-completion tests. Bormuth interpreted cloze readability scores as follows:

Cloze Scores	Reading Level
0% to 34%	Frustration Level
35% to 49%	Instructional Level
50% and above	Independence Level

Scores below 35% indicate an inability to gain "information" from the passage. Scores between 35% and 49% indicate an ability to gain information with instructional assistance. Scores beyond 50% represent an ability to gain information from texts independently.

In short, more than twenty years of research has established the validity, reliability, and utility of the cloze procedure as a tool for estimating readability. Bormuth's work on standards of reliability, however, should make it evident that there is only a tenuous distinction between the cloze as a test of readability and the cloze as a test of comprehension. Bormuth's study on readability, as a matter of fact, is often cited in discussions of the validity of the cloze as a test of reading comprehension, and his "standards of readability" are used to interpret cloze comprehension scores (e.g., Hansen and Hesse, 1974). A cloze readability score tells a teacher something about the characteristics of the text in relation to the reading competency of the students, with the emphasis, as the term "readability" implies, on the text. As a test of comprehension, the cloze procedure generally remains identical, but the interpretation shifts from characteristics of the text to characteristics of the student.

The tenuous distinction between readability (comprehensibility) and comprehension, however, is not peculiar to the cloze; rather it is inherent in the concepts themselves. Readability formulas, for example, are usually validated with standardized reading lessons in comprehension as a criterion. In the Lorge, Flesch, and Dale-Chall formulas, the criterion is the Standard Test Lessons in Reading (McCall and Crabbs, 1925, 1950, 1961).

When the cloze procedure, rather than a standard, multiple-choice test of comprehension, is used as a criterion, readability formulas "consistently yield higher predictive validity coefficients" (Klare, 1974, p. 66). This implies that the cloze procedure has more in common with readability formulas than standardized comprehension measures. Bormuth (1971), on the other hand, suggests that cloze tests measure an even broader range of skills than traditional, multiple-choice comprehension tests. But that may be a disadvantage.

Scores on traditional comprehension tests already reflect such a broad range of psycholinguistic skills that it is nearly impossible to specify exactly what the tests measure.

Comprehension

Syntactic cues. A student attempting to replace missing words in connected discourse has two basic decisions to make: (1) He must decide which part of speech is appropriate to the syntactic context and (2) which particular word within that grammatical category is appropriate to the semantic context. The student makes both decisions on the basis of his knowledge of the syntactic and semantic regularity of the language.[4] If the sentence is within the grammatical competence of the student, he has enough syntactic cues (i.e., the order in which the morphemes occur) to choose the appropriate part of speech even though he may not know what the content words mean. A nonsense sentence, retaining only the morphemes (underlined) necessary to parse the sentence, makes the distinction between syntactic and semantic decisions clear: "*The* ______________ lca scuokk*ed* tcon*ly* *down* *the* eezbu rgag." The missing word obviously performs an adjectival function in the sentence. Thus a student faced with a gap in the following sentence--"The ______________ car careened madly down the canyon road"--has enough syntactic cues to know that the missing word again has to behave like an adjective. In grammatically well formed English sentences, that is, determiners like "the" are usually[5] followed by nouns, adjectives, or adverbs (e.g., "the happily soused man") while verbs are usually preceded by noun phrases (nouns or pronouns, and modifiers).[6]

---

[4] The assumptions are that the original sentence is grammatically well formed, that the words are part of the lexicon of the language, and that the particular combination of words "makes sense" to other members of the speech community.

[5] Note an exception in this sentence: "the" is followed by an auxiliary verb, "are."

At the very least, then, the cloze procedure assesses the student's syntactic competence, and that competence is fundamental to the comprehension of any sentence. "To comprehend a sentence, the reader must understand the underlying structural relationships, i.e., the logical subject and logical object of the sentence" (Simons, 1970, p. 33; Fodor and Garrett, 1967; Fodor, Garrett, and Bever, 1968; Weisberg, 1971; Smith, 1975).[7]

Not only is syntactic competence fundamental to comprehension, but the apprehension of structural relationships in a sentence is considered part of the process of comprehension since syntactic and semantic processes are intimately bound up with each other in language performance.[8] It is impossible, that is, to "assign meaning to words in a sentence without knowing how the words are grouped which implies...[a knowledge of] the syntactic structure of the sentence" (Miller, 1965, p. 17). The apprehension of meaning clearly includes grammatical relationships if "meaning" is construed

> as the total disposition to make use of and react to a linguistic form. It follows that a readiness to use words in accordance with conventions about the parts of speech is a part of meaning. However, it is a part that can be distinguished from reference. (Brown, 1958, p. 118)[9]

---

[6] See Chomsky (1957, 1965) for an analysis of syntactic structures.

[7] There is some experimental evidence "that perception, comprehension, and recall of sentences is intimately connected with underlying sentences" (Finn, 1973). See studies by Lenneberg (1967), Anderson (1973), and Fodor and Bever (1965).

[8] Note that syntactic theory (a competence/knowledge theory) was originally developed without recourse to semantic theory, but the discussion here concerns the use of various competencies.

[9] The referential aspect of meaning is only peripheral to the apprehension of meaning in connected discourse where "the interanimation of words" (Richards, 1936/1967) predominates. "The meaning of an utterance is not a linear sum of the meanings of the words that comprise it" (Miller, 1965, p. 18).

Identifying the apprehension of the meaning residing in structural relationships as an essential act of comprehension as such has several advantages over traditional, global conceptualizations of comprehension: (1) It enables researchers to distinguish clearly between more rudimentary reading skills, like decoding (recognition of letters as sound and groups of letters as words), or word knowledge, which may be prerequisite to comprehension. (2) It also enables researchers to specify the relationship between the conceptualization of comprehension and actual linguistic components.

There is some empirical evidence that cloze scores reflect syntactic competence to a greater degree than traditional reading comprehension tests. Simons (1970), for example, devised a "Deep Structure Retrieval Test" (D.S.R.T.). Students were asked to identify the anomalous sentence among three sentences, two of which were paraphrases of each other. Scores were then correlated with cloze scores and scores on the Metropolitan Achievement Test (M.A.T.).

> The correlations between the D.S.R.T. and the Cloze Test are significant and quite large, with more than 50% of the variance accounted for by the D.S.R.T. The relationship between the D.S.R.T. and the M.A.T. Reading is significant but not as great as the Cloze Test. (p. 74)

Simons concluded that

> Recovering the deep structure is an important aspect of reading comprehension. In fact Ss' skill at recovering the deep structure of sentences is a much more important aspect of reading comprehension skill, as measured by a cloze test, than I.Q., word knowledge and word recognition skill. (p. 89)

Semantic cues. The cloze procedure, on the other hand, has been criticized as a test of comprehension on the assumption that "cloze scores are probably more dependent on detection of grammatical than of semantic cues" (Carroll, 1972, p. 19). As Ramanauskas (1972) points out, however,

> It is difficult to separate semantic and syntactic sources of constraint experimentally although they can be distinguished conceptually. Brown (1970) for example, wrote that syntactic expectancies are guided by prior semantic information, as in the search for a logical subject and predicate. (p. 324)

Much more research needs to be done on the relationship between cloze scores and the syntactic and semantic components of language. It does seem clear, nonetheless, that grammatical cues only allow the student to pick the appropriate part of speech for a missing word in a cloze passage. Exact-word replacements (or synonyms), on the other hand, require an apprehension of the semantic cues surrounding the missing words. A student "must guess what the mutilated sentence means as a whole, then complete its pattern to fit that whole pattern" (Taylor, 1953, p. 416).

Guessing missing words in context is not far removed from the actual process of reading connected discourse. The student's

> habits of reading cause him to anticipate words, almost automatically, when he is receiving messages. When he sees the start of a phrase that looks familiar, he immediately tends to complete it in his own way even when the written phrase actually ends differently. (Taylor, 1953, p. 419)

Goodman (1970) describes such reading habits as a "psycholinguistic guessing game":

> Efficient reading does not result from precise perception and identification of all elements, but from skill in selecting the fewest, most productive cues necessary to produce guesses which are right the first time. The ability to anticipate that which has not been seen, of course, is...vital in reading, just as the ability to anticipate what has not yet been heard is vital in listening. (p. 260)

Readers can guess missing words in connected discourse not only because of the syntactic regularity of the language but also because there is considerable semantic redundancy in any utterance.

> "Man coming" means the same as "A man is coming this way now." The latter, which is more like ordinary English, is redundant; it indicates the singular number of the subject three times (by "a," "man," and "is"), the present tense twice ("is coming" and "now"), and the direction of action twice ("coming" and "this way"). Such repetitions of meaning, such internal ties between words, make it possible to replace "is," "this," "way," or "now," should any of them be missed. (Taylor, 1953, p. 418)

Carroll (1966), distinguishing between concepts and words, ascribes the semantic redundancy in "normal language texts" to the overlap of the concepts (verbal and non-verbal classes of experience) "suggested by the words in a sentence" (p. 84).

The recurrence of particular expressions in a speech community also increases the probability of certain words occurring in specific sentences. Taylor (1953), notes, for instance, that "'Please pass the ______________' is more often completed by 'salt' than by 'sodium chloride' or 'blowtorch'" (p. 419). The probabilities obviously vary with the situational context. For example, "salt" might occur more often in that sentence at the dinner table, but "sodium chloride" might be more frequent in the chemistry lab or "blowtorch" in the welding shop. Ordinarily, in connected discourse, the sentence would be embedded among other sentences, further defining the semantic context and constraining the number of words that would be appropriate.

Taylor's examples, however, are mostly cliché expressions grounded in social "rituals." Though all "semantic regularity" is ultimately based upon shared experience (verbal and non-verbal),[10] the cloze procedure is no less effective when dealing with sentences that a reader has probably never encountered before. Consider, for example, the "mutilated" sentence introduced earlier in this discussion of the cloze procedure: "The ______________ car careened madly down the canyon road." The reader has enough syntactic cues to know that the missing word has to behave like an adjective, but the list of appropriate words in this semantic context excludes many adjectives from consideration (assuming that the rest of the words in the sentence are within the reader's vocabulary and can be related to the reader's non-verbal experience). Cars careening madly down canyon roads, for instance, are not likely to be "supercilious" even though "supercilious" can perform the functions of an adjective as required by the grammatical context. Embedding the sentence in a cohesive paragraph would further reduce the number of adjectival expressions that would be appropriate in this sentence.

---

[10] "Shared experience" implies psycho- and sociolinguistic systems only dimly understood at present. See Bernstein (1969).

On the other hand, cloze tests often contain missing words which are very difficult to replace no matter how extensive the context (Fletcher, 1959; Bormuth, 1962). Who could guess the deleted word in the following sentence, for example, without knowing the original text? "I then took up three planks from the flooring of the chamber, and deposited all between the ____________" (Poe, "The Telltale Heart"). (The missing word is "scantlings.")

The occurrence of both easy and difficult restorations does not present insolvable problems for the cloze procedure.

> A series of about 50 blanks is roughly sufficient to allow the chances of mechanically selecting easy or hard words to cancel out and yield a stable score of the difficulty of a passage, or the performance of an individual, despite what specific words the counting-out process may delete. (Taylor, 1956, p. 48)[11]

In addition, Bormuth (1967a) contends that very easy and very difficult restorations of deleted words contribute to a test's validity "in testing subjects differing widely in ability" (p. 12).

Contextual constraint. There is limited empirical evidence regarding the extent of contextual constraint on cloze deletions. MacGinitie (1961) varied the deletion patterns on two prose narrative passages and randomly assigned 20 college students "to each omission set of each passage." "No statistically significant difference was found in the difficulty of restoring omitted words when every 24th, 12th, or 6th word was omitted, but omitting every 3rd word made restoration more difficult" (p. 125). MacGinitie concluded that "additional uninterrrupted context beyond five words did not help in the restoration of the missing word" (p. 127). (Note that MacGinitie's results are only based on two passages, both 144 words long, that his subjects are college students, and that the ten word bilateral constraint is only an average over the two passages. No attempt was made to identify contextual clues[12] or to relate the contextual constraint of specific deletions to meaningful units of discourse, e.g., independent clauses.) Aborn, Rubenstein, and Sterling (1959) also found that a context of five to ten words was maximally effective in the replacement of missing words. Their study, however, is based upon isolated sentences rather than connected discourse. Taylor (1956) also reports that every-fifth-word deletion is "statistically independent" in cloze tests.[13]

---

[11] Deleting every fifth word 50 times would require a passage of at least 250 words. Most cloze research, following Taylor, is based on passages of approximately 250 words in length (Potter, 1968).

Other studies of contextual constraint compared unilateral and bilateral constraint (context preceding or following and context surrounding omissions). Weaver (1962) discovered "that a context is most restrictive when a word is embedded within it. Bilateral context seems to improve the precision of language" (p. 155). Indeed, Coleman and Miller (1968) found "that the bilateral constraint is so great that surprisingly little information is added to it by reading the passage" (p. 374).

---

[12] Ames (1966) attempted to identify contextual clues, and Rankin and Overholser (1969) investigated "the sensitivity of intermediate grade pupils to contextual clues described by Ames" (p. 50).

[13] "It should be noted that cloze materials for first graders have been modified to make it possible for them to cope with this type of task" (Rankin, 1974, p. 6). Gallant (1965), for instance, had to use a three-option, multiple-choice cloze to maintain test reliability in grades 1, 2, and 3. Gove (1975) used passages of less than 75 words in length and deleted only lexical items.

Reducing the context for interpretation to five or ten words implies that "cloze scores are dependent chiefly on what might be called the 'local redundancy' of a passage, i.e., the extent to which linguistic cues in the immediate environment (generally, in the same sentence) of a missing word tend to supply it" (Carroll, 1972, p. 18). Thus MacGinitie's evidence seems to run counter to Taylor's contention (1953) that a student "must guess what the sentence means as a whole, then complete its pattern to fit that whole meaning" (p. 416). Further, a context for interpretation of five to ten words excludes larger semantic units, "the major ideas or concepts that run through a discourse" (Carroll, 1972, p. 19), whereas "it is typical and natural for sentences to be comprehended as part of a larger semantic unit" (Dooling, 1972, p. 56). Moreover, "comprehending a sentence in context is a more complex task than comprehending a sentence in isolation" (p. 60). Any test of comprehension, therefore, must get at larger units of meaning than five to ten word clusters.

Qualifying his generalization about contextual constraint in the cloze, MacGinitie (1961) writes that

> Although it seems that constraints between words generally decrease very rapidly with distance, this does not mean that constraints never operate over distances of more than four or five words. Also, some constraints, such as knowing the topic of the paragraph, may have a more generalized influence that does not decline with decreasing length of context in an easily specifiable way. Carroll, Carton, and Wilds (1959) report that when a paragraph is broken into 10-word segments with the 5th word in each segment omitted, restoration is much less accurate when the segments are presented in random order rather than in their original order. (p. 128)

Ramanauskas (1972) also gathered evidence on intrasentential constraints by assigning two cloze tasks to educable, mentally retarded students. One task presented students with "selections containing sentences in the natural order of discourse...[and] the other task involved materials wherein the sentence order was modified by being randomly rearranged" (p. 338). Ramanauskas found that "a significantly greater number of correct cloze responses were produced for material having sentences in the natural order of discourse" (p. 342). Moreover, Fillenbaum (in Potter, 1968), found that while "form class predictability is more dependent upon the immediate grammatical environment...verbatim predictability depends upon both this factor and remote topical content or semantic features of the discourse" (p. 23). Thus there seems to be both logical and empirical evidence for the cloze as a measure of both small and larger semantic units in connected discourse, i.e., a measure of reading comprehension.

Correlation and factor analytic studies. In addition to this kind of logical and piecemeal empirical evidence, many investigators have studied the relationships between cloze scores and scores on standardized, norm-referenced tests of reading comprehension. The sample of these investigations displayed in Table 3.1 indicates, in general, a substantial correlation between such scores. Moreover, Rankin (1965) notes that, with few exceptions,

> comparisons between cloze tests and standardized reading tests have yielded substantial correlations even though the cloze tests were based upon a variety of different types of reading materials and were constructed and administered in different ways. (p. 136)[14]

Since standardized, norm-referenced tests of reading comprehension are biased toward critical reading skills, it's not surprising that cloze scores, with a significant syntactic factor, correlate substantially rather than highly with scores on standardized comprehension tests, as indicated in Table 3.1. Nor is it surprising to discover "that correlations with cloze scores are frequently higher for vocabulary measures than for comprehension measures" (Potter, 1968, p. 5), as illustrated by Rankin (1957) and Fletcher's (1959) studies in Table 3.1. With a significant syntactic factor and a preponderance of intrasentential constraint, cloze scores should correlate highly with measures of comprehension between the polar extremes represented by pre-comprehension vocabulary measures and tests of comprehension biased toward critical reading skills.[15]

---

[14] Such variations, however, make it difficult to compare results from different studies. Construct validation of the cloze as a test of comprehension becomes even more difficult.

In any event, the kinds of correlations represented in Table 3.1 are often accepted as indicative of the validity of the cloze procedure as a test of "general comprehension," as it is usually called in the literature, or, more specifically, the ability to comprehend. Moreover, as a test of comprehension ability, the cloze has few of the liabilities of standardized, norm-referenced tests. Test construction in the cloze procedure, for instance, requires no particular expertise in language or testing and is sufficiently objective (even "mechanical") to allow for the construction of parallel test forms for periodic testing. More importantly, there are no questions in the cloze procedure to introduce extraneous difficulties and processes. Cloze tests are, however, cumbersome to grade since they have to be scored by hand.

There is some conflicting evidence regarding the cloze procedure as a test of ability in reading comprehension. Weaver and Kingston's (1963) study, for example, as indicated in Table 3.1, is an exception to the general tendency toward substantial correlations between cloze scores and scores on standardized comprehension tests. After examining "the relationships of cloze tests to standard tests of reading, listening and language symbolizing ability," they concluded that the "cloze tests are related only moderately to the verbal comprehension factor" (p. 259).

---

[15] Which is exactly what does happen. See the discussion of "specific comprehension" on pages 23 and 24.

Table 3.1

Correlations Between Cloze Readability Tests and Standardized Tests of Reading Achievement

Study	Subjects	Tests	Correlations
Jenkinson (1957)	High School	Cooperative Reading C2	
		Vocabulary	.78
		Level of Comprehension	.73
Rankin (1957)	College	Diagnostic Survey	
		Story Comprehension	.29
		Vocabulary	.68
		Paragraph	.60
Fletcher (1959)	College	Cooperative Reading C2	
		Vocabulary	.63
		Level of Comprehension	.55
		Speed of Comprehension	.57
		Dvorak-Van Wagenen	
		Rate of Comprehension	.59
Hafner (1963)	College	Michigan Vocabulary Profile	.56
Ruddell (1963)	Elementary	Stanford Achievement Paragraph Meaning	.61-.74
Weaver & Kingston (1963)	College	Davis Reading	.21-.51
Gallant (1965)	Elementary	Metropolitan Reading	.65-.81
Greene (1965)	College	Diagnostic Reading Survey Total Comprehension	.51
Heitzman & Bloomer (1967)		Iowa Reading	.26-.68
		Differential Aptitude Verbal Reasoning	.33-.86
Geyer & Carey (1972)	Jr. High	Standardized Reading Test	.53

Bormuth (1969) questioned Weaver and Kingston's interpretation of the data "on at least four counts": (1) Their subjects were a highly select group of college students; (2) "the correlations upon which they based their calculations differed in size from those obtained by other investigators"; (3) "the standardized tests they used showed unusual patterns of factor loading"; and (4) "the cloze tests showed some inconsistencies among themselves in their loading patterns" (p. 361).

Bormuth then set out to investigate further the factor validity of cloze tests. Nine passages of approximately 250 words each were clozed, and seven multiple-choice tests written on each of the passages.

> The [multiple-choice] tests were written to measure comprehension of vocabulary, of explicitly stated facts, of sequences of events, of stated causal relationships, of the main ideas of the passages, of inferences, and of the author's purpose....An equal number of each type of item was written for each passage....The items were then administered to samples of subjects enrolled in grades four, five, and six. (p. 361)

Bormuth found that "the intercorrelations were high and fairly uniform across the different types of tests" (p. 363), and concluded that "clearly one factor accounted for the preponderance of the variance. Further, there was little difficulty in applying the name of 'reading comprehension ability' to that factor" (p. 364).

Though Bormuth's study is an important contribution to cloze research, labeling the factor upon which both types of tests loaded reading comprehension "ability" is an unfortunate misnomer that obscures important distinctions. Standardized comprehension tests are refined, highly developed tests of general verbal ability. Except for the opinions of three "reading specialists," Bormuth made no effort to validate his multiple-choice comprehension tests. Bormuth's criterion test, in contradistinction with standard comprehension tests, is weighted toward "literal comprehension" (e.g., 108 vocabulary items, 63 items dealing with explicitly stated facts, and only 36 "inferential" items). Furthermore, Bormuth compared cloze and comprehension test scores on the same passages, resulting in a measure of what Rankin (1965) calls "specific comprehension" or comprehension *per se* as distinguished from Weaver and Kingston's (1963) attempt to measure "general comprehension" or "general verbal ability."

Several other investigators have constructed multiple-choice and sentence-completion comprehension tests in order to compare cloze scores with comprehension scores on the same passages and thereby to evaluate the cloze procedure as a measure of "specific comprehension." Correlations between cloze test results and comprehension scores on the same passages are generally high as would be expected in light of the preceding discussion. Taylor (1957) got a correlation of .80; Jenkinson (1957) .82; Friedman (1964) .90 to .91; and Bormuth (1962) .73 to .84 (or .93 when cloze tests and comprehension test results were combined first and then correlated).

> In summary, the cloze procedure appears to be a highly valid measure of the specific comprehension of a particular message. In fact, it is a more accurate measure of specific comprehension than of general reading skill as measured by standardized reading tests. (Rankin, 1965, p. 136)

It should again be noted, however, that the criterion tests used in these kinds of studies of the cloze procedure are seldom validated (Potter, 1968).

Information gain. The cloze procedure has also been used to measure "information gain" (sometimes referred to as "knowledge," "reading," or "learning gain"). Information gain is assessed by testing comprehension before and after reading the passage upon which the test is based, and then taking the difference between the two scores as a measure of information gain.[16] In an attempt to test "learning" (information) gain with the cloze procedure, Taylor (1957) constructed five different test forms based upon a long (3,240 words) technical article. As criterion measures, Taylor used "two matched comprehension tests," one designed to measure pre-test knowledge of the material in the article and the other to assess knowledge of it immediately after study. He then constructed three forms of a pre-reading cloze test and three forms of a post-reading cloze test on a 20% sample of the same article.[17] The type of deletion varied from any-word, to "hard" words (nouns, verbs, and adjectives), to "easy" words (verb auxiliaries, conjunctions, pronouns, and articles) on the three forms. Students were allowed to study the article immediately before attempting to restore missing words on the post-reading cloze. Taylor concluded that "'any' and 'hard' [cloze tests] yielded equally significant learning gains, ones somewhat larger than the corresponding comprehension tests did" (p. 26). In order to select a representative sample of the 3,240 word article for the cloze tests, however, Taylor mechanically selected eight nine-line subsamples, for a total of 650 words, and artificially joined them together. The results are therefore suspect as a measure of the comprehension of connected discourse.

Rankin (1957, 1959) also attempted to measure knowledge gain with the cloze procedure and found the most significant gain scores with a modified cloze procedure. "The correlation between the pre-cloze, noun-verb deletion test and the criterion test was .86 (corrected for attenuation)" (Rankin, 1974, p. 137).

---

[16] The kind of "information" gained obviously depends on the kinds of comprehension questions, so there is no more specificity inherent in the use of "information gain" than there is in "comprehension." Subtracting pre- from post- reading test scores does, however, allow the investigator to reduce the measure of pre-test knowledge in test scores. If comprehension is the assimilation of information in the text to cognitive structures in the reader, such a reduction is absurd.

[17] Usually referred to as "pre-cloze" and "post-cloze" tests in the literature. The standard cloze procedure results in a pre-cloze test.

Coleman and Miller (1968), using the standard, any-word deletion, found that "the cloze score before reading...is measuring essentially the same information as the cloze score after reading. The correlation between the two was .93" (p. 374). They concluded that "the bilateral constraint is so great that surprisingly little information is added to it by reading the passage" (p. 374). Greene (1964) also found little difference between pre- and post-reading cloze scores when deleting any word. More research is needed, but modifications of the cloze procedure seem to be more viable as tests of information gain (Rankin, 1974).

Interpreting Cloze Scores. Several investigators have attempted to develop standards for interpreting cloze scores. Bormuth (1967b) determined that a score of 38% correct restorations on a conventional cloze test[18] is equivalent to 75% on a specially constructed, multiple-choice comprehension test on the same passage (a test of "specific comprehension" as defined above). If the multiple-choice score is corrected for guessing, the equivalent cloze score is 43%. Since the cloze test was a measure of "specific" rather than "general comprehension" and since the multiple-choice test was not validated against any established comprehension test, the results of this study cannot be generalized. In 1968, however, Bormuth used the California Achievement Test as a criterion measure and found that cloze scores of 44% and 57% were comparable to reading achievement test scores of 75% and 95% respectively. Rankin and Culhane (1969) came within an average 3.1 percentage points of replicating Bormuth's 1967 results, with greater differences toward the extremes, "particularly toward higher multiple-choice percentage scores" (p. 197), which they attributed to ceiling effects on Bormuth's multiple-choice test. Table 3.2 indicates the comparable scores in all three studies.

---

[18] A "conventional" or "standard" cloze test is defined as a cloze test where passages are 250 words or more in length, every fifth word is deleted, only exact-word replacements are scored as correct (minor misspellings excepted), and the test is given under untimed conditions. Most cloze research has conformed to these strictures (Potter, 1968; Rankin, 1974), hence the label, "conventional" or "standard."

Table 3.2

Cloze Test Percentage Scores Comparable to 75% and 90% Criterion Multiple-Choice Scores

	Comparable Cloze Percentages		
Criteria	Bormuth (1967)	Bormuth (1968)	Rankin & Culhane
75%	38	44	41
90%	50	57	61

In light of the fact that two of the three multiple-choice tests were unvalidated, that the inconsistencies in the conception and nature of comprehension between the three multiple-choice tests affect comparable cloze scores rather strongly as indicated in Table 3.2, Rankin and Culhane's (1969) conclusion that "it is now possible for teachers to interpret cloze scores with some degree of confidence by using specific percentage points as criteria of acceptable performance" (pp. 197-198) seems overstated.

In the most thorough study of cloze criterion scores to date, Bormuth (1971) explored the relationships between conventional cloze tests and various criterion measures, including "measures of information gain, rate of reading, willingness to study, and preferences for the subject matter, style, and level of difficulty" (p. viii). Cloze scores and multiple-choice or sentence-completion scores were compared on identical passages. "Comprehension" was conceptualized as "information gain," the difference between pre- and post-reading scores on the multiple-choice or sentence-completion tests. The notion of comprehension was further restricted to "the information explicitly signaled" in the passages (p. 21) or to "what is commonly called literal comprehension" (p. 117).[19] Though Bormuth's model is admittedly incomplete and tentative (p. 20), the results are generally consistent with previous studies. The relationships between cloze scores and information-gain scores varied considerably from grade to grade as illustrated in Figure 3.1. Bormuth interpreted cloze scores between 35% and 49% as indicative of the appropriateness of the text for instructional uses, and cloze scores between 50% and 70% as indicative of independence level textual material.

Bormuth's study, as mentioned previously, is primarily concerned with "standards of readability," but Hansen and Hesse (1974) used these criterion scores to interpret comprehension test scores in Madison public schools, and the results were unexpected. Large proportions of the students seemed to be reading below the literacy level as defined by Bormuth's criterion scores. It should be noted, however, that Hansen and Hesse (with Bormuth as consultant) used cloze passages of less than standard length (60 to 70 words) whereas Bormuth's criterion scores were developed with passages of 250 or more words. Moreover, Bormuth's criterion measure in the 1971 study was again unvalidated. Much more work needs to be done on cloze criterion scores, and, until firmly established, cloze scores must be interpreted cautiously. Finally, research on cloze criterion scores has been limited to the standard cloze procedure and is inapplicable to variations in format and type of deletions.

---

[19] On the basis of this assertion alone, Bormuth's contention that cloze tests involve a broader range of skills than those "normally identified and measured in multiple-choice comprehension tests...[including] those that are so complex and difficult that they fall above the upper limits of the multiple-choice tests" (p. 32) is misleading. The notion of "information gain" in this study is hardly comparable to the critical reading skills assessed by standardized, multiple-choice comprehension tests.

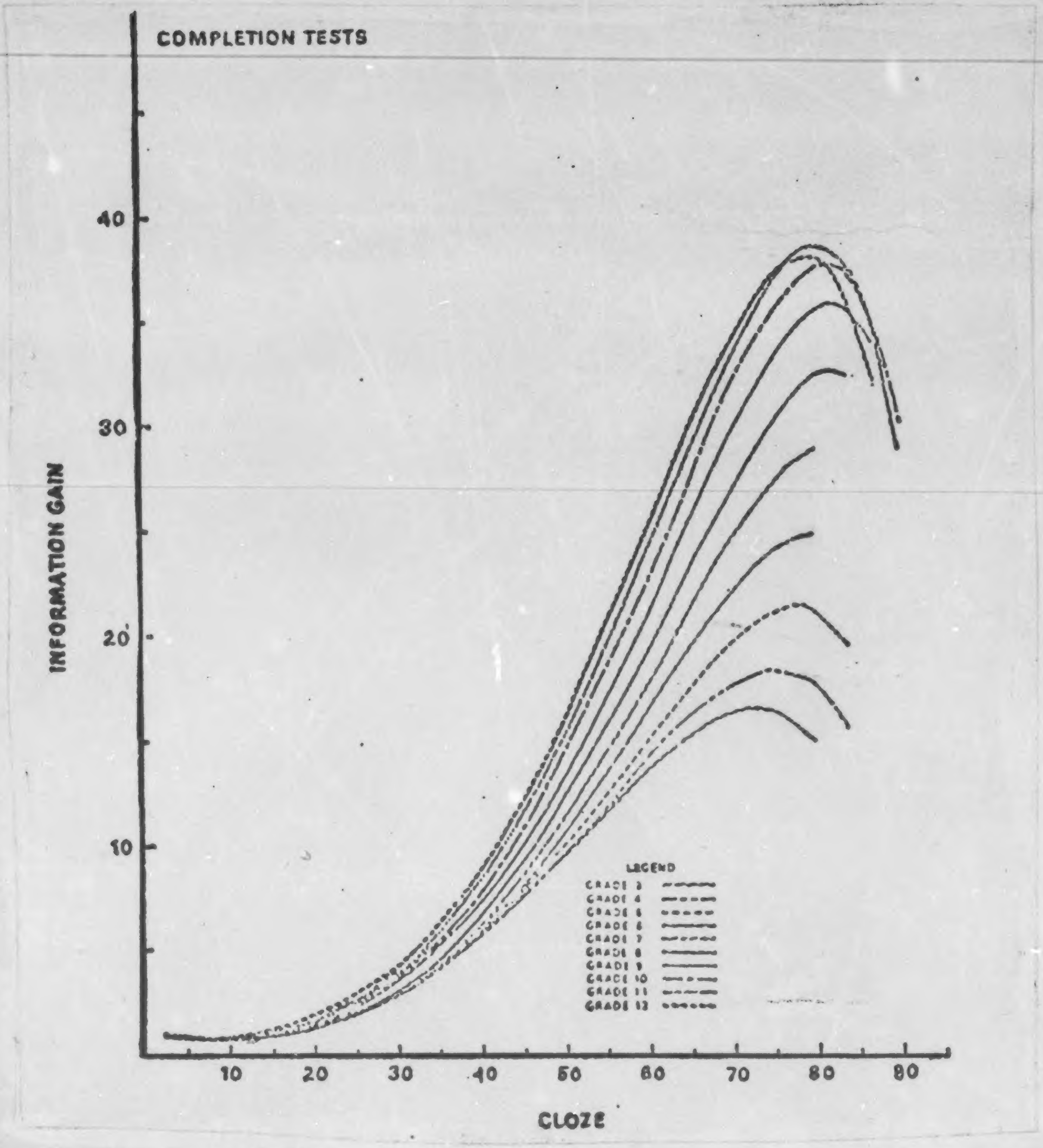


Figure 3.1. Regression of information gain scores on cloze scores. (From Bormuth, J.R., Development of Standards of Readability: Toward A Rational Criterion of Passage Performance. Final Report. Chicago: University of Chicago, June 1971, p. 102. [ERIC DOCUMENT ED 054 233].)

Deletion types. Though most researchers have used the standard, any-word deletion procedure when investigating the cloze, there is some evidence that the deletion of particular parts of speech (e.g., nouns) may be used to separate comprehension *per se* from general verbal abilities. In the discussion of "information gain," for instance, it was noted that the deletion of nouns, verbs, and modifiers ("lexical" or "content" words as they are commonly called in the literature)[20] seems to produce better reading gain scores than conventional, any-word deletions. Further, Rankin (1974) maintains that "the almost exclusive reliance upon . . . [the standard] cloze has strengthened the influence of general verbal abilities and intelligence upon the cloze measurement of reading comprehension" (p. 3). (Many investigators [Taylor, 1953; Rankin, 1959; Fletcher, 1959; Deutsch et al., 1974; Ruddell, 1965; Schneyer, 1965] have found substantial correlations between cloze scores and measures of intelligence, especially general verbal ability.) Taylor (1953) and Rankin (1959) compared any-word deletions with "lexical" deletions and found that correlations between cloze scores and IQ were diminished by deleting only "lexical" words.

In a more elaborate effort to explore the relationships between types of deletions and comprehension scores, Louthan (1965) constructed seven different types of cloze tests from each of 24 prose passages, 500 to 600 words long, using a 10% deletion ratio,[21] and administered the tests to 236 seventh-grade pupils. The seven kinds of deletions were any-word, nouns, verbs, modifiers, prepositions and conjunctions, determiners, and pronouns. A control group got passages with no deletions. The pupils attempted to replace the missing words on the cloze passages and then answered 12 "comprehension"questions (six "factual" and six "inferential") on the same passages without referring back to the passages.[22] Louthan found that the deletion of "lexical" words significantly affected student ability to answer comprehension questions and concluded that nouns, verbs, and modifiers "are the basic meaning carriers of the written language" (p. 297).

---

[20] Any-word deletions, on the other hand, are commonly called "structural" deletions on the assumption "that the total amount of structural meaning in a passage would be reduced more than the total amount of lexical meaning, if cloze tests were constructed by deleting every '*n*th word'" (Rankin, 1974, p. 4). "Structural deletion" was an unfortunate phrase, however, and only created confusion (Rankin, 1974).

[21] 20% is "standard".

Several other investigators have analyzed deletion types in the cloze procedure. Ohnmacht et al. (1970), for example, studied "the relationships of flexibility of closure and speed of closure to a number of cloze tasks representing structural [any-word], lexical, concrete, and abstract deletions" (p. 206). Like Louthan, Ohnmacht et al. concluded that "a lexical deletion is considered to sample a different construct, 'comprehension,' because nouns, verbs, and adjectives seem to have a good deal to do with such 'comprehension' components as vocabulary" (p. 215). Bickley, Weaver, and Ford (1968) also investigated the effect of deletion by grammatical categories. When nouns, main verbs, adjectives, and function words[23] were deleted separately, "only the deletion of nouns had a significant effect on S's ability to supply multiple-choice answers" (p. 614). Moreover, referring to a previous study,[24] Bickley et al. note that "the most deleterious condition was the blacking out of nouns, main verbs, and adjectives [simultaneously] leaving mostly the function word categories; Ss who had no reading paragraphs at all did better than these Ss" (p. 614). Bradley (1970) extended the findings by Bickley et al. to "lower grade and age levels" (p. 92).

---

[22] It should be noted that the directions for this test administration strengthen the memory factor in comprehension (see Carroll, 1972), necessarily making it more difficult to answer questions when "content" words are missing.

[23] "All words blacked out except nouns, main verbs, and adjectives" (p. 613).

[24] Weaver and Bickley, 1967.

In general, then, it seems that restricting deletions to so-called lexical words in the cloze procedure reduces the syntactic while heightening the semantic component in cloze scores. (It is neither possible nor desirable to eliminate the syntactic component since it is part of the process of comprehension. Lexical words obviously carry structural as well as semantic information.) In light of the small number of studies devoted to deletion types, however, such an interpretation of the findings is tentative, especially considering the vague and various notions of comprehension evidenced in the multiple-choice, criterion measures. Indeed, as Ohnmacht et al. (1970) remark,

> The fact that responses to cloze tasks reflecting essentially gross deletion strategies align themselves with crude measures of comprehension does little to shed light upon the fundamental nature of comprehension other than to indicate that one can measure what passes for comprehension in more than one way....Researchers using the cloze procedure ought to give careful consideration to language operations and to rational operations which are implicit in verbal activity and they should construct deletion patterns which seem to relate to these operations. Rather than standardizing a particular cloze deletion type, exploration of a wider range of deletion types which are related to particular linguistic and psychological hypotheses is needed. (pp. 215-216)

Summary and Conclusion

The cloze procedure was developed as a "new tool for measuring readability" in 1953, and more than 20 years of research since then has firmly established the cloze as a measure of readability, comprehension, and related areas of inquiry. As a measure of readability, the cloze has proven far more accurate than readability formulas, but the comparison is misleading. Any tool that measures is likely to be more accurate than one that predicts.

The accuracy of measurement, however, in no wise satisfies the need for predictability. As a matter of fact, most passages on cloze comprehension tests are graded for readability with the Dale-Chall formula (Potter, 1968). As a test of readability, the cloze procedure sacrifices the convenience of predictability for the accuracy of measurement.

But it is pointless to compare the cloze procedure with readability formulas. The cloze is an unorthodox comprehension test which is so easy to construct that it can be used to determine the comprehensibility (readability) of any text for a particular group of students in short order. Rather than comparing the cloze procedure with readability formulas, it would be more meaningful to compare the cloze with the traditional comprehension tests used as criterion measures in developing readability formulas. Readability formulas, for example, yield consistently higher predictive validity coefficients when the cloze is used as a criterion measure, and this implies that the cloze procedure is more accurate as a test of reading comprehension than traditional comprehension tests. The emphasis in cloze research has gradually shifted from readability to comprehension.

As a measure of reading comprehension, the cloze has several advantages over traditional, multiple-choice comprehension tests. As noted previously, cloze tests are easy to construct, requiring no particular expertise in language or testing. Moreover, cloze tests sample the syntactic and semantic content of a passage more objectively and thoroughly than any other comprehension test. Most importantly, the cloze procedure does not disrupt the process of comprehension with extraneous difficulties and processes in the form of questions which are sometimes more difficult to comprehend than the passage itself.

On the contrary, guessing missing words in connected discourse seems to be very similar to the way skilled readers actually read. Reading, that is, appears to be a "psycholinguistic guessing game." Readers make decisions about the interpretation of the interrelationships among the words of the discourse based upon the information they cull from the syntactic and semantic cues in the text and their previous verbal and non-verbal experience.[25] These interpretations in turn create expectations in the mind of the reader for congruous information.[26] The reader then "reads ahead," "predicts" words and groups of words he has not yet read on the basis of his expectations about the text. The cloze procedure, rather than disrupting this process with extraneous skills, only slows it down, forces the reader's attention to the linguistic information in the text that partially governs his interpretive decisions, the resulting expectations of congruous information, and the predictions about the parts of the text not yet seen. The cloze procedure, that is, forces the reader to focus on the syntactic and semantic context surrounding the missing words. Thus the attention is where it belongs in the comprehension of printed discourse--on the "interanimation of words" in the text and on the integrative faculties of the reader.

---

[25] The reader's guesses are also predicted on the basis of his own syntactic and semantic "system." There is evidence that readers store the text as paraphrase, i.e., that they process text for meaning in terms of their own syntactic and semantic system (Lenneberg, 1967). If there is little or no congruence between the syntactic and semantic systems of the text and those of the reader, the text is of course incomprehensible to the reader.

[26] Good writers, of course, "play" with these expectations, delaying, momentarily thwarting them, surprising the reader with a more inclusive resolution than he had expected. This kind of "play" goes on within sentences as well as in larger, thematic units.

In the process of emphasizing context, the cloze procedure may tend to focus attention on smaller syntactic and semantic units than would ordinarily be the case in perusing a text for meaning. A few studies, based upon a limited number of passages, indicate that most of the syntactic and semantic information needed to supply missing words in connected discourse comes from the six to ten words surrounding each deletion.[27] The cloze procedure, that is, seems to force conscious interpretation of units of discourse that a skillful reader would ordinarily subsume (or "recode" or "chunk" as the process is sometimes labeled) in larger units of meaning. That is, "in comprehending sentences in discourse, Ss construct a 'theme' or 'schema'; they reduce the information into larger semantic units" (Dooling, 1972, p. 60). Forcing the reader to consciously interpret smaller semantic units, however, does not in itself indicate an insensitivity to larger semantic units (both "explicit" and "implicit")[28] which bind the sentences of a discourse into a larger unity of meaning. As a matter of fact, disrupting the normal order of the discourse causes cloze scores to fall off, and this surely indicates a sensitivity to context beyond the six to ten words surrounding the missing word. The extent of that sensitivity, however, is still an open question.

Additional evidence for the validity of the cloze as a comprehension test comes from correlational studies. Correlations between cloze scores and scores on standardized, norm-referenced tests of reading comprehension are generally substantial, and this is indicative of the validity of the cloze as a measure of general comprehension or the ability to comprehend the kinds of reading materials sampled by the test.

---

[27] Whether or not this is a general feature of written English remains to be determined. The interplay between immediate and remote context can vary greatly from text to text, and the effect of this variation on contextual constraint has not been studied.

[28] e.g., themes or tone. A particular restoration might be appropriate to its immediate syntactic and semantic context, but violate the tone of the message as a whole. (When students, queried about their guesses, respond that it "sounded right," their explanation may not be so superficial as it sometimes seems.)

That such correlations are substantial rather than high is consistent with the foregoing analysis. Standardized, norm-referenced tests of reading comprehension are biased toward a conceptualization to reading as reasoning and emphasize general verbal ability at the expense of more specific and fundamental comprehension skills. Cloze scores, on the other hand, have a strong syntactic factor. Moreover, as noted above, there seems to be a predominance of intrasentential constraint governing the restoration of missing words in cloze tests. These two observations suggest that the cloze procedure measures a lower level of comprehension than standardized, norm-referenced tests. The tendency of cloze scores to correlate higher with the vocabulary than with the comprehension sections of standardized tests provides more evidence of a similar sort. The standard cloze procedure thus seems to measure a level of comprehension somewhere between the polar extremes represented by vocabulary and comprehension scores on standardized tests.

Findings from studies of the cloze as a test of specific comprehension or comprehension *per se* (as distinguished from general verbal ability) also imply that cloze restorations make fewer demands on the reasoning powers of the reader than items on standardized comprehension tests. Correlations between cloze scores and multiple-choice comprehension scores on the same passages are consistently high. Moreover, an examination of test items on these specially constructed, multiple-choice comprehension tests indicates a bias toward "literal comprehension" (because such items are easier to write and replicate?). Thus the cloze is usually considered more valid as a test of specific than of general comprehension and seems to get at more fundamental comprehension skills than standardized, norm-referenced tests.

Finally, selective deletions (e.g., verbs only) in the cloze procedure raise the possibility of identifying and manipulating the role specific linguistic components and form classes play in the comprehension of connected discourse. Conceptualizations of comprehension could then be stated in such a fashion as to lead to testable hypotheses and empirical investigation. There is already some evidence that comprehension _per se_ can be extricated further from general verbal abilities and reasoning processes by limiting deletions to lexical words (nouns, verbs, adjectives, and adverbs). Deleting only lexical words also seems to reduce the syntactic while heightening the semantic factor in cloze scores. The multiple-choice cloze testing system described in the next section of this report is a further extension of this line of inquiry.

CHAPTER IV

SPPED[1] CLOZE EXERCISES IN A MULTIPLE-CHOICE FORMAT

Twenty years of cloze research has neither produced an entirely satisfactory cloze comprehension test nor has it silenced the critics of the cloze. While the standard cloze procedure has some decided advantages over traditional comprehension tests, it also has some serious liabilities. Most cloze research is based on the any-word, every-fifth-word deletion pattern, and the almost exclusive use of such a pattern seems to have resulted in a measure that loads too heavily on syntax and general verbal ability (Rankin, 1974). In addition, the free-response cloze, where the student writes in the missing word, is not amenable to machine scoring and makes a horrendous task out of scoring tests, particularly with large numbers of students. The modified cloze procedure discussed in this chapter is an attempt to respond to these and other criticisms of the cloze procedure as a test of comprehension.

The Standard Cloze Procedure

As defined in the preceding chapter, the "standard" or "conventional" cloze procedure is a mechanical technique for deleting every fifth word in a text of at least 250 words and replacing the deleted words with underlined blanks of a standard size. Students who have not been allowed to read the original text are then asked to write in the missing words with no other clues to their identity than the mutilated text. There are no time constraints on the task, and only exact replacements are counted as correct.

---

[1] System for Pupil and Program Evaluation and Development.

Advantages

Question-free item type. The standard cloze procedure has several obvious advantages over traditional reading comprehension tests of the passage and question type. Most commentators point out the ease and objectivity of constructing a cloze comprehension test--there are no questions to write and no sets of distractors to produce; indeed, given a passage, there are no subjective decisions of any kind to make. Considering the theoretical and practical difficulties posed by the construction of traditional comprehension test items, that is no mean advantage. It was argued in the first two chapters of this proposal, for instance, that (1) it is incumbent upon the test-maker to specify and control the relationship between text and test items else there is little possibility of determining what the test measures, and (2) traditional test items, including the wh-item, either control the relationship between item type and text to the exclusion of the semantic component of the discourse, or introduce the test-writer's own idiosyncratic interpretation of the text into the test items, thereby sacrificing objectivity and passage dependency. In addition, the test items themselves often introduce comprehension difficulties which are extraneous to the test passages. If the cloze procedure can avoid such problems and still produce a viable test of comprehension, then it is a boon to test-makers and teachers alike.

Comparable to the reading process. Another major advantage of the cloze procedure, less often mentioned in the literature, is that a cloze comprehension test elicits decisions from the student which are very similar to those decisions students ordinarily make in attempting to comprehend printed discourse. Reading is a constructive language process; that is, any reader has to reconstruct the meaning intended by the writer from minimal information on the printed page. The interaction between the reader and the text might be described in brief as follows: The reader comes to any segment of connected discourse with expectations about what that discourse means based upon the verbal and non-verbal context in which it occurs. Likewise, the reader comes to any sentence in the discourse with expectations about what it means based upon his apprehension of the meaning of previous sentences and the verbal and non-verbal context in which they appear. Given the expectation of meaning of a particular sort, the reader then searches for the logical subject and logical predicate of the sentence which will fulfill that expectation. In effect, the reader makes an hypothesis about what a sentence means, and then samples among the linguistic clues to meaning in the text in an effort to verify his hypothesis. (An "objective" reader is a fiction; readers are always biased --that is, selective in their perceptions.) If the hypothesis is quickly verified, the sampling procedure can be very attenuated. If, on the contrary, the hypothesis is not immediately verified (and readers can be remarkably blind to contradictory information), the reader may sample more extensively in an attempt to verify his original hypothesis or may change his hypothesis and sample again. And so on.

Now the demands made by the cloze procedure on a student are not far removed from the psycholinguistic processes implied by this model of reading as a constructive language process. The cloze procedure does not unduly disrupt the reading process. That is to say, a competent reader is always "reading ahead," making predictions about what a given sentence should mean, and then sampling the linguistic clues to meaning in the sentence in order to verify his predictions. The cloze procedure, in a comparable manner, asks the student to predict the meaning and identity of words in a discourse based upon the student's apprehension of the meaning of previous segments of the discourse and the verbal and non-verbal context in which they occur (i.e., the test directions, the testing situation, etc.). The possibility of reconstructing the surface structure of the discourse exactly as the writer had intended it even though 20% of the words are missing is made possible by the natural redundancy of language, the well-formedness of the discourse, and the shared psycho- and sociolinguistic systems of reader and writer. These are, of course, requisite conditions for the comprehension of any discourse, deletions or no, so the cloze procedure requires no peculiarly redundant texts.

Specificity. In addition to ease and objectivity of test construction and a general similarity in the demands that both the reading process and the cloze procedure make upon readers, the cloze procedure also makes it possible to identify and control the interaction between text and item type (deletion type and rate). Analysis of deletion rates, for instance, indicates that most of the syntactic and semantic information needed to replace a missing word is found in the six to ten words surrounding the deletion. More precisely, the information needed to replace function words (e.g., prepositions, determiners) is generally found in closer proximity to the deletion than the information needed to replace "content" words (e.g., nouns, verbs) (Fillenbaum, Jones, and Rapoport, 1963). Thus the extent of the verbal context within which interpretive decisions are made can be specified in the cloze procedure.

Moreover, the relative influence of syntactic and semantic clues in the text on cloze scores can be determined. In the standard cloze procedure, for instance, the any-word, every-fifth-word deletion pattern produces a preponderance of the syntactic component of the text in cloze scores simply because most sentences have a greater proportion of syntactic than semantic clues. Observations of this kind have led Rankin (1959, 1974) to dub the standard cloze "the structural cloze." Correlations with measures of the comprehension of syntactic structures (Simons, 1970; Stedman III, 1971) also indicate a strong syntactic factor in cloze scores. Since an apprehension of the syntactic structure of a sentence is fundamental to its comprehension, the standard cloze procedure measures a more basic, identifiable level of comprehension than standardized comprehension tests.

Disadvantages

Local redundancy.

What is perceived as an advantage from one perspective, however, can just as readily be characterized as a disadvantage from another. Carroll (1972), for example, has criticized the standard cloze for its dependence on syntactic cues and insensitivity to the train of ideas that runs through a discourse and binds it together. Brown (1970) also identifies the standard cloze as a more rudimentary measure than comprehension--assimilation to cognitive categories. Though it is impossible to separate syntax from comprehension, in general, it does seem to be true that the standard, any-word, every-fifth-word deletion pattern produces a measure of comprehension unduly weighted toward syntax and thus unduly dependent on local redundancy.

Exact-word-only.

In comparing the cloze procedure to a model of reading as a constructive language process, it was noted above that the standard cloze procedure required a student to predict not only the meanings of words but their specific identities. The difference between predicting meaning and exact-word-only replacements of missing words marks a clear line of demarcation between the standard cloze and the model of reading as a constructive language process. While the reader ordinarily tries to reconstruct the meaning of a written message as represented by the orthographic system on the printed page, he may do so in terms of his own syntactic and semantic structures. That is, when asked to recall the meaning of an utterance, a student will often reply in terms of his own competence; he reconstructs the meaning in his own language patterns. Slobin and Welsh (1967), for example, cite the following exchange between a model and a two-and-one-half-year-old child:

> Model: This one is the giant, but this one is little.
>
> Child: dis one little, annat one big. (p. 8)

(Fillenbaum [1970] cautions, however, that memory and comprehension are easily confused in such analyses.) Moreover, the sampling procedure of the model of reading as a constructive language process also implies a rough match (rather than exact replication) between the surface structure represented in the orthography on the printed page and the reconstructed message in the mind of the reader. Finally, it was posited in the second chapter of this proposal that the surface structure of language is never more than an approximation of the meaning intended by the writer or the meaning apprehended by the reader.

The standard cloze procedure, on the other hand, demands that a reader not only reconstruct the meaning of a message from the clues in the text, but that he reconstruct exactly the same orthographical representation of the meaning intended by the writer. Now that is clearly demanding something in addition to "comprehension." Rankin (1974), as a matter of fact, has cautioned teachers against trying to justify exact replacements when using the cloze procedure as a teaching device. Goodman (cited in Fiske, 1975) has also warned teachers against "correcting" student approximations of a text while reading. Standard cloze scores therefore seem to indicate something more than a student's apprehension of meaning in connected discourse.

Taylor (1953), who brought the cloze procedure to the attention of the reading field, was quite clear about the relationship between cloze scores and the apprehension of meaning: The percentage of correct responses on a standard cloze test indicates "the extent of likeness between the language patterns used by the writer to express what he meant and those possibly different patterns which represent readers' guesses at what they think the writer meant" (p. 417). The match between the language patterns of the writer and reader is more demanding than comprehension normally makes upon a reader. As a consequence, cloze scores are usually quite low in comparison to scores on traditional comprehension tests on the same passages.

Passage length. Moreover, the great range in the difficulty of replacing individual words on a clozed passage makes it necessary to use passages of 250 or more words so that the test score reflects a measure of the average difficulty of the passage. But passages of that length make domain-referenced testing difficult--few representative passages could be used in the time available in any testing period. Furthermore, a great (haphazard) range in difficulty indicates that the test-maker is incapable of specifying exactly what the test measures since the interaction between text and deletion is not sufficiently specified and controlled. While the standard cloze procedure has obvious advantages over traditional comprehension tests, it still amounts to another global measure of "comprehension," whatever that is.

Hand scoring. In the standard cloze procedure, students have to write in the missing words, and test administrators must score each test laboriously by hand. Nothing so reduces the utility of the cloze as the necessity of hand scoring. Until a viable cloze procedure is developed in the multiple-choice format, the cloze procedure will be relegated to use in small classrooms only.

The Modified Cloze Format of the SPPED Cloze Exercises

The discussion of standardized, norm-referenced reading comprehension tests, wh-items, and the standard any-word deletion type of the conventional cloze procedure should make it evident that none of these item types produces a satisfactory test of literal comprehension. Among the three item types discussed, however, the cloze procedure clearly offers the best possibility for objectively and thoroughly sampling the student's apprehension of "the grammatical and semantic relations which obtain within and among the sentences of the discourse."

Moreover, the cloze procedure offers the test-maker the opportunity to identify and control the interaction between characteristics of the student, the text, the item type, and the testing situation. Any test of reading comprehension should identify and control the interaction of such characteristics in order to specify what the test actually measures, but the need for an explicit construct becomes particularly acute when an attempt is made to label a test or subsections of it according to the "level" of comprehension (e.g., literal, inferential) it attempts to assess. The state of psycholingustic knowledge, however, allows for nothing more than a first, tentative effort to identify and control such interacting characteristics. What follows, then, is (1) a brief, condensed, and tentative statement of a construct of literal comprehension based upon the evidence and analyses adduced in preceding chapters of this proposal, (2) a rationale for the modified cloze format adopted in the SPPED cloze exercises based upon that construct, and (3) a description of the actual construction of those exercises.

A Tentative Construct of Literal Comprehension

Since reading comprehension was defined in Chapter II as the apprehension of the meaning(s) of written discourse, it is evident that comprehension is a successful synthesis of the competence of the student with the demands made upon that competence by characteristics of the written discourse in question. In the testing situation, the particular demands made upon the student's competence are primarily controlled by the test items. Contingent circumstances external to the student which affect a correct synthesis of features of the text accessed by the test items and the competence of the student are considered part of the test situation in this analysis. In addition, the dispositional limitations of the student can affect either the attainment of the requisite competence or the use of such competence in the testing situation.

Characteristics of the student. It is assumed by the construct that students who can comprehend at the literal level have gone through the normal stages of cognitive and linguistic development appropriate to their chronological age group. More specifically, it is assumed that these students have no physical or psychological impairments that hinder normal language development, reading ability, or test performance. IQ's of these students are assumed to be 85 or higher. If any of these assumptions is violated, then students may not perform as predicted below.

It is hypothesized that comprehension at the literal level demands the same competencies and no other competencies on the part of the student at any grade level. Two of these competencies are: a general knowledge of standard-English-speaking societies[2] and a basic competence with the English language.[3] Basic linguistic competence subsumes (a) lexical knowledge, (b) a semantic rule system for selecting appropriate senses from the meanings of lexical items, and (c) a syntactic rule system for interrelating selected lexical senses. A third competency is the ability to recognize and differentiate between different orthographic representations of different lexical items and their sequential appearences in unique combinations in print as sentences of the English language.

---

[2] A distinction is made here between language- and culture-specific competencies. Speakers of standard English as a second language, for example, may have the linguistic competence to comprehend many texts at the literal level, but culture-specific knowledge, with which writers usually assume readers are acquainted and which they therefore fail to state explicitly, may, on occasion, thoroughly confound the non-native speaker's efforts to comprehend literally a given message.

Characteristics of the text. The literal comprehension of connected discourse assumes that (1) the text in question is in fact connected discourse. The reading materials must be grammatically and semantically well-formed sentences in standard English that pursue a particular topic, situation, description, idea, etc., coherently; that is, there is no willy-nilly introduction of new topics, ideas, etc., from sentence to sentence in the text. In addition, it is assumed that the text has in fact a literal level to be comprehended; that is, that the discourse is not so excessively idiomatic, metaphorical, or esoteric as to confound deliberately any effort to comprehend it at the literal level. If either of the above assumptions is violated, then students may not perform as predicted below.

If the student has the competencies outlined above and the text does not make excessive demands upon those competencies, then the student, properly motivated and given the opportunity and conditions to do so, will comprehend the text at the literal level. That is, the student will apprehend "the grammatical and semantic relations which obtain within and among the sentences" of the text. If the student fails to comprehend literally, then the text has exceeded his psycholinguistic competency. That is, the syntactic structures in the text are too complex for the student to "parse," and/or the vocabulary and concepts in the text surpass his lexical knowledge.

---

[3] The construct is stated in terms of the English language but could easily be restated in terms of any other language or in general terms.

Characteristics of the test items. The apprehension of the literal meaning of connected discourse is defined as the apprehension of "the grammatical and semantic relations which obtain within and among the sentences of the discourse." That is, the context for interpretation is limited, insofar as possible, to the sentences of the discourse itself. Such a definition of literal comprehension is considered the lowest possible synthesis of the linguistic components of the text with the psycholinguistic competency of the student that can be labeled comprehension of connected discourse without violating conventional understanding of "comprehension" or "connected discourse." Test items which purport to measure literal comprehension, therefore, must access "the grammatical and semantic relations which obtain within and among the sentences of the discourse" and only those relations. Passage dependency then, is essential to literal comprehension test items. That is, if the grammatical and semantic information necessary to the selection of the correct response from among a set of responses is present in or implied by the set of responses or other unspecified contextual features, then the correct response cannot be cited as evidence of literal comprehension. Moreover, test items must sample objectively and adequately among the grammatical and semantic relations of a discourse, or test scores may not be cited as evidence of the literal comprehension of that discourse. For example, if the item type subordinates the semantic component of the discourse to its syntactic component, then correct responses to such an item type may not be presented as evidence of literal comprehension.

If, however, the item type accesses only the grammatical and semantic relations of the discourse, and does so thoroughly and objectively, then correct responses to such item types can be adduced as evidence of literal comprehension and no other "level" or "degree" of comprehension. That is, if literal comprehension is defined as the lowest possible synthesis of the linguistic components of the whole text and the psycholinguistic competence of the student, and if the test items only access such a synthesis, then these test items will measure literal comprehension or no comprehension and nothing else.

Characteristics of the testing situation. All those criteria that normally apply in testing situations in order to elicit the best possible performance from the student--e.g., appropriate time of day, a minimum of distractions, etc.--are assumed by the construct, literal comprehension. Two additional assumptions are stressed: (1) Students must be familiar with the item type. Unconventional item types obviously require training, or students may not perform as predicted. (2) Students must have sufficient time to work carefully through all the test passages. Students who are rushed through passages on a test may not perform as predicted. Other than the motivational factor, the test situation should be essentially neutral to the construct of literal comprehension.

Rationale for the Modified Cloze Format Used in the SPPED Cloze Exercises

Deletion type. Only nouns, verbs, adjectives, and adverbs are deleted[4] on the assumption that (1) such words carry most of the information which is unique to any given discourse. Function words (determiners, prepositions, auxiliary verbs, etc.) certainly convey information too, but such information is mostly structural; that is, function words primarily define the interrelationships between the appropriate senses of the lexical items of the discourse. Moreover, though syntactic analysis is vital to the comprehension of any verbal message, the information communicated by syntactic analysis alone is not unique to a particular message. Syntax rarely tells the reader anything he does not already know (Katz, 1972).

---

[4] Only nouns and verbs are deleted in grade 1 and 2 materials.

(2) Research cited in Chapter III also indicates that the deletion of nouns, verbs, and modifiers reduces correlations between cloze scores and measures of general verbal ability or IQ while simultaneously increasing correlations with measures of "specific comprehension" and "information gain." Substantial or high correlations with IQ are antithetical to the construct, literal comprehension. In addition, criterion measures used in studies of "specifc comprehension" and "information gain" tend toward literal comprehension in conceptualization.

(3) Fillenbaum et al. (1963) found that the grammatical and semantic information necessary to replace nouns, verbs, and modifiers in connected discourse tended to be further removed from the deletion than the information needed to replace function words. Topical content, for instance, which is more likely to affect the particular choice of nouns, verbs, or modifiers in a given text, is liable to be dispersed throughout the text. It is assumed, therefore, that the restriction of the deletion type to nouns, verbs, and modifiers will obviate some of the criticism of the cloze as a measure of local redundancy and increase the semantic component of the text in cloze scores without eliminating the syntactic component, Nouns, verbs, and modifiers also convey structural information, albeit to a lesser degree than function words.

(4) Finally, deleting only nouns, verbs, adjectives, and adverbs makes it feasible to construct a cloze test in the multiple-choice format. That is, relatively few words can function as determiners or auxiliary verbs in an English sentence; making up distractors for "the," for instance, seems pointless.

On the other hand, the number of words that can function as nouns in an English sentence is enormous. In addition, deleting nouns, verbs, and modifiers makes it possible to select distractors that are specific to content areas (e.g., social studies), a prerequisite to viable distractors in a domain-referenced test.

Deletion rate. An every-fifth-word deletion rate (20% of the text) is considered optimum because it samples the grammatical and semantic relations of the text objectively and as thoroughly as possible without depriving the student of the information he needs to replace the deleted words. Selective deletion types (e.g., nouns), however, force some variation in deletion rate. For example, the test-maker who is trying to maintain an every-fifth-word deletion rate while deleting only nouns, verbs, and modifiers, often counts five words in a text and finds no candidate for deletion. If he backs up below three words between deletions, it becomes very difficult to replace missing words with so little remaining, immediate context. On the other hand, if he counts too far in the other direction, then the thoroughness of the sample of the grammatical and semantic relations begins to suffer. A test-maker cannot readily reject passages on the grounds that they create difficulties for the adopted deletion pattern, else the passages will represent a biased selection from the domain of reading materials. Consequently, the every-fifth-word deletion rate is adhered to as much as possible; there are never fewer than three words between deletions; occasionally there are as many as eleven words between deletions.

Distractors. All distractors are (1) grammatically plausible and (2) semantically implausible. (1) Grammatical plausibility means that any distractor can perform properly the grammatical function assigned to it by the syntactic position of the missing word for which it functions as a distractor. For example, if a noun is deleted, the distractors are usually "nouns" or words that can behave like nouns in the syntactic position in question. Distractors are grammatically plausible in order to reduce further the syntactic component in test scores since grammatical meaning, as such, does not account for very much of the information which is unique to the text in question.

Since the distractors are all grammatically plausible, syntactic analysis provides no definitive basis for choosing among the sets of responses. On the other hand, syntactic analysis is fundamental to the comprehension of the interrelationships among the words of any sentence, and no test of comprehension should or could short-circuit that analysis. Consequently, the sets of responses associated with each deletion are located below each passage, encouraging the student to process the grammatical and semantic information in the text, to make an initial hypothesis about the meaning and identity of the missing word before looking for verification of that hypothesis among the sets of responses. In other words, an attempt is made to preserve the best feature of the standard cloze procedure--its similarity to the reading process--while introducing a multiple-choice format. In summary, syntax is present in the modified cloze format adopted in the SPPED cloze exercises, but it is not preponderant in cloze scores.

(2) In the set of responses associated with any deletion, only the correct answer, the exact word in the original text, is semantically plausible. Again, the attempt is made to preserve the best features of the standard cloze procedure while weeding out its liabilities. It was noted in the first section of this chapter, for instance, that requiring students to supply exact replicas of missing words in the standard cloze procedure is antithetical to the model of reading as a constructive language process. On the other hand, in that search for the basis of the commonality of meaning intended by the

writer and apprehended by the reader which is literal comprehension, it was posited that the common meeting ground between writers and readers is the orthographic representation of meaning on the printed page, and therefore that representation is the closest possible approximation of what the writer meant. Accordingly, the modified cloze procedure maintains the insistence on exact-word-only replacements. The student who makes a correct hypothesis about the meaning of a missing word based upon his apprehension of "the grammatical and semantic relations which obtain within and among the sentences of the discourse" will have no difficulty modifying his hypothesis about the surface representation of that meaning when confronted with the correct answer among a set of distractors. The distractors only behave like traditional distractors when the vocabulary level or the syntactic complexity of the passage exceeds the student's competence, that is, when the student can no longer comprehend at the literal level.

No attempt is made to tamper with the approximation of the meaning intended by the author. No attempt is made to interpret the text for the student, to impose the test-writer's own idiosyncratic interpretation of the text on the text in the form of distractors that compete with the correct response. Such semantic competition, such alternate possibilities for interpretation, are antithetical to the construct, literal comprehension, which is rooted in commonality of interpretation rather than nuances of meaning. Cranney (1972) found, for instance, that semantically plausible distractors in a multiple-choice cloze format introduce significant numbers of items into the cloze test that are even more difficult than the hardest items on standard cloze tests. Semantically plausible distractors extend the context for interpretation beyond the grammatical and semantic relations of the discourse which is, again, antithetical to the construct, literal comprehension. It is hypothesized that semantically plausible distractors will also maintain or increase correlations between the cloze procedure and measures of general verbal ability or IQ. Such correlations are also antithetical to the construct, literal comprehension. Finally, semantically plausible distractors, with an emphasis on nuances of meaning, make the cloze procedure into a very difficult vocabulary test. Literal comprehension does not demand so extensive or refined a vocabulary.

The Construction of the SPPED Cloze Exercises

The construction of the SPPED Cloze Exercises was undertaken in order to test the efficacy of the multiple-choice cloze format as a measure of literal comprehension and as a means of readily implementing, with reproducible test items, the concept of domain-referenced testing. At the outset, a plan was devised for the systematic sampling of reading materials in four domains in which students are expected or required to read. The domains are:

1. Textual Material in Reading/Literature, Language Arts, Social Studies, Science and Mathematics;
2. Citizen Material (newspapers and news magazines);
3. Consumer Material (catalogs, advertising, instructions, and so forth); and
4. Reference Material (test instructions, children's magazines, encyclopedias, and so forth).

Selection of cloze passages. Textual materials were to be sampled at each grade level, from 1 through 10. Materials in the other domains were to be assigned to grade levels on the basis of readability scores. Quotas were established for the number of samples to be collected, at random, for each grade and domain. The resources used for the sample collection were the New York State Education Department's Curriculum Laboratory and the State Library. In addition, some consumer passages were taken from

(Hansen and Hesse, 1972).

The selection procedures resulted in the identification of 1,374[4] passages that were coherent and of specified lengths appropriate for clozing. Their distribution by domain and grade level is shown in Table 4.1. Table 4.1 also shows the distribution of the textual materials by subject matter.

Determination of readability. All of the passages--those in the textual domain as well as those in the citizen, consumer, and reference domains--were subjected to readability calculations so that they could be ordered by difficulty for test construction purposes. The readability formulas used were the Spache (1953, 1960) and the Dale-Chall (1948). (As noted previously, the cloze itself has advantages over conventional formulas as a measure of readability. However, the Spache and Dale-Chall are widely used measures, and their utility as rough indices of difficulty is borne out by the results of the initial use of the cloze passages reported in Chapters VIII and IX.)

The Spache is normally used for grades 1 through 3, the Dale-Chall for grades 4 through 12 and college. Both formulas use average sentence length and percent of "hard words" in calculating difficulty. "Hard words" are those not appearing on lists of familiar words. The word list for the Spache formula is "Clarence Stone's Revision of the Dale List of 796 Easy Words"; for the Dale-Chall formula it is the "Dale List of 3,000 Familiar Words." (The criteria for difficulty used in devising both formulas were graded reading materials.) The Spache formula produces grade level scores. The Dale-Chall formula produces raw scores interpreted as "corrected grade levels." The corrected grade level for a raw score of 5.0 to 5.9 on the Dale-Chall, for example, is fifth to sixth grade.

---

[4] Another 120 passages have been added to the citizen domain and expansion of the textual domain into college levels is anticipated.

Table 4.1

Passages for Cloze Testing

Grade	Domain									Grand Total
	Textual						Citizen	Consumer	Reference	
	Reading	Lang. Arts	Math	Sci.	Soc. St.	Total				
1	48				30	78				78
2	41				30	71				71
3	30	20	20	20	20	110			10	120
4	42	20	20	20	20	122			9	131
5	36	20	20	20	20	116	3		10	129
6	33	20	20	20	20	113	6	8	12	139
7	30	20	20	20	20	110	6	9	11	136
8	30	20	20	20	20	110	5	9	10	134
9	30	20	20	20	20	110	16	15	10	151
10	34	20	20	20	20	114	20	11	10	155
11							20	13	8	41
12							19	13	10	42
13							14	12		26
14							11	10		21
Total	354	160	160	160	220	1054	120	100	100	1374

The range of Spache scores was divided into six equal intervals, and the range of Dale-Chall scores was divided into 22 equal intervals. This gave 28 difficulty levels covering grades 1 through college. The raw scores, difficulty levels, and original grade level interpretations given by Dale-Chall and Spache are shown in Table 4.2.

Use of the readability formulas disclosed wide ranges of difficulty among instructional materials at given grade levels. Both extremely easy and extremely difficult passages that differed markedly from other materials for the same grade were eliminated in the selection process. However, there is still variation in the number of difficulty levels covered by grade levels in the textual domain. The grade level of the source was used as the guide in application of the cloze procedure. Both grade level and difficulty level are indicated by the identification number for each passage.

Preparation of cloze items.[5] The procedure for word deletion in the cloze passages varied with the grade of the source. In grade 1 and 2 materials, every eighth word was deleted, and deletions were limited to nouns and verbs. For grade 3 and above, every fifth word was deleted. Deletions included adjectives and adverbs as well as nouns and verbs.

In all cases, the initial deletion was made between the sixth and tenth words. The exact starting point was determined by a table of random numbers. The number of deletions per passage was fixed by the passage length, which varied by grade level. The number of alternatives in the multiple-choice responses also varied by grade level: three alternatives

---

[5] Only the briefest summary of the modified cloze procedure is given here. See Appendix A for a complete description of the passage selection and item-writing procedures.

Table 4.2

Difficulty Levels for Cloze Passages

Readability formula	Raw score	Difficulty level	Original grade level assignments by Spache and Dale-Chall
SPACHE	1.0-1.4	1	1
	1.5-1.9	2	
	2.0-2.4	3	2
	2.5-2.9	4	
	3.0-3.4	5	3
	3.5-3.9	6	
DALE'CHALL	4.50-4.74	7	4
	4.75-5.99	8	
	5.00-5.24	9	5-6
	5.25-5.49	10	
	5.50-5.74	11	
	5.75-5.99	12	
	6.00-6.24	13	7-8
	6.25-6.49	14	
	6.50-6.74	15	
	6.75-6.99	16	
	7.00-7.24	17	9-10
	7.25-7.49	18	
	7.50-7.74	19	
	7.75-7.99	20	
	8.00-8.24	21	11-12
	8.25-8.49	22	
	8.50-8.74	23	
	8.75-8.99	24	
	9.00-9.24	25	13-15 (College)
	9.25-9.49	26	
	9.50-9.74	27	
	9.75-9.99	28	

at grade 1, four at grades 2 and 3, and five at grades 4 and above. These variations by grade level are summarized in Table 4.3, Specifications for Cloze Passages and Test Items.

The correct multiple-choice response to a cloze item is the exact word deleted from the passage. To assure distractors of appropriate difficulty for the test items, graded lists of nouns, verbs, adjectives, and adverbs were prepared using Harris and Jacobson's Basic Elementary Reading Vocabularies (1972) and EDL Research and Information Bulletin 5: A Revised Core Vocabulary (Taylor, Frackenpohl, and White, 1969). Special content words for subject matter areas like Social Studies were compiled using the Harris-Jacobson material and the American Heritage Word Frequency Book (Carroll, Davies, and Richman, 1971).

Initially, distractors were selected from appropriate lists by use of a table of random numbers. Later, a computer program was written for automatic random selection of distractors. Each set of distractors was reviewed to eliminate tricky or ineffective distractors, such as synonyms, and to assure that the distractors agreed with the stem in tense, number, and so forth.

With a minimum of 3 deletions per passage at grade 1 and a maximum of 10 deletions per passage at grade 3 and above, nearly 15,000 multiple-choice items have been prepared for the SPPED Cloze Exercises.

Format. All cloze passages and test items were put in a comparable format. The format gives (1) the identification number of the passage, (2) a title (provided by the item writer), (3) the passage itself, and (4) the test items. Large (Bulletin) type was used for the first two grades. A sample cloze passage for grade 2 is shown in Figure 4.1.

Table 4.3

Specifications for Cloze Passages and Test Items

	Grade 1	Grade 2	Grade 3	Grade 4 and above
Passage length	25-35 words	40-45 words	60-70 words	60-70 words
Words deleted	Nouns Verbs	Nouns Verbs	Nouns Adjectives Verbs Adverbs	Nouns Adjectives Verbs Adverbs
Frequency of deletions	Every 8th word	Every 8th word	Every 5th word	Every 5th word
Deletions per passage	3	5	10	10
Alternatives per item	3	4	4	5

02-02-01-01-01-001

WHAT DOES ANDY SEE?

Andy saw something at his bedroom ____1____! He ran as fast as he could to ____2____ Mother.

"Listen, Mother!" said Andy.

Mother said, "Please ____3____, Andy. I have to get ready to ____4____ to work now. Mrs. Coats is ____5____ for me."

① a. paper
b. window
c. apple
d. oven

② a. wash
b. tell
c. buy
d. fix

③ a. peep
b. fly
c. point
d. wait

④ a. think
b. sing
c. tip
d. go

⑤ a. finding
b. roping
c. waiting
d. racing

Figure 4.1. Sample cloze passage and items.

CHAPTER V

APPLICATION OF THE MULTIPLE-CHOICE CLOZE IN MEASUREMENT AND EVALUATION

The multiple-choice cloze materials are one component of a testing system intended to offer to the educational community more useful and adaptable measures of reading comprehension than are currently provided by standardized tests. This chapter will first briefly describe that testing system, the Test Development Notebook (TDN). It will then point out various advantages and features of the multiple-choice cloze materials, and discuss the utility of these materials for a variety of evaluation and decision-making purposes. Next, the chapter will present the principles involved in applying the multiple-choice cloze materials in specific testing situations. The chapter will conclude with a description of the test assembly and administration procedures followed in the first experimental application of the multiple-choice cloze passages and items.

The Test Development Notebook

The initial chapter of this report cited rigidity of format as one of the major shortcomings of standardized reading tests. The Test Development Notebook was originally conceived as a flexible test-construction resource that would provide school districts with large numbers of reading items, identified by different skills or objectives and difficulty, which could be assembled in different ways to meet different evaluation needs. The TDN was at first planned to include several different formats which might measure unique aspects of comprehension. To date there are sizable item pools for two of these item formats, the multiple-choice cloze and the wh-/main idea items.

The multiple-choice cloze item pool, the development of which was described in detail in the previous chapter, consists of approximately 1,374 clozed passages (i.e., generally, 60-70 word passages with ten deletions and accompanying multiple-choice items) categorized (temporarily) by readability levels determined by Spache and Dale-Chall readability formulas. The wh-/main idea item pool consists of 300 passages, 15 at each of 20 readability levels, whose lengths vary systematically by readability level (e.g., approximately 25 words at level 1 and up to 220 words at levels 17-20). Each of these passages is accompanied by up to four multiple-choice main idea items and up to eight multiple-choice wh-detail items modeled after Bormuth's (1970) wh-items. The formats of the cloze and the wh-materials are both generative procedures for preparing numbers of parallel, multiple-choice items.

The concept of the TDN, which is currently a "paper bank," derives from the computerized and paper-based approaches to test assembly formalized in such projects as the Sequoia Comprehensive Achievement Monitoring (CAM) program in Redwood City, California, where some 60,000 items have been banked to support local test assembly. The organization of the TDN, however, has also benefited from several years of experience in developing and refining CAM in schools in New York State. Much was learned from the New York CAM experience about the practical aspects of making some of the newer concepts in evaluation work broadly in practice. (Referred to here are several years of applying complex evaluation designs, such as longitudinal matrix sampling, that are now routinely used in schools as part of local CAM projects.) In addition, the TDN, particularly in the design specifications for the multiple-choice cloze materials, draws upon Hively's (1974) model of domain-referenced testing. The organization of the TDN considerably improves on the Sequoia project and similar efforts. Instead of simply filing and providing for accession of items and related information, the TDN is ultimately generative for both item and test production. That is, the cloze item format appears to be capable of conversion to an algorithm which can be used to process any appropriate sample of written discourse into items. An indefinite number of such items can, therefore, be generated. In addition, finished items in the TDN can be accessed and organized into an infinite number of tests by a process that interfaces directly with the actual printing and production of test forms.

The TDN is, therefore, an attempt to build a generalized test assembly resource. At its current stage of development, the multiple-choice cloze component of the TDN is the most important aspect of this generalized test assembly resource, for the multiple-choice cloze materials permit measurement of literal comprehension across the total range of interest, in a variety of evaluation contexts (e.g., tests may be assembled to assess a first grader's ability to comprehend literally a basal reader or a high school student's ability to comprehend literally texts in the content areas). The following sections describe the properties of the multiple-choice cloze materials, and accompanying advantages thereof, which are outlined in Table 5.1

Table 5.1

Properties and Advantages of Multiple-Choice Cloze Testing Materials

Property	Advantage
Objectivity in item format	The format of the multiple-choice cloze adheres to the concept of an item form, an objective and generative procedure for producing items that are an unbiased representation of a universe of content. The application of this principle here avoids the problems of subjectivity and resultant content bias in test construction procedures, criticisms which have been characteristically leveled at tests of reading comprehension.
Domain-referenced content	Items or passages have been systematically sampled from recognized content areas representing relevant domains of written discourse. Test scores can ultimately be generalized to a domain(s) of written discourse with specified properties of readability, content, etc., thus making reading test scores more directly useful in reading instruction.
Unidimensionality	The multiple-choice cloze format is a unitary and generalizable measure of reading comprehension, appropriately termed "literal comprehension." Literal comprehension is the reading behavior most affected by the instructional program.
Equal-interval scaling	Cloze passages are to be calibrated on an equal-interval scale. This represents a substantial improvement over existing scaling procedures in tests of reading comprehension. Passages in the test will be referenced to meaningful upper and lower limits. Any test assembled from the item pool will be referenced to this single scale which in turn can be related meaningfully to objective performance criteria.
Passage dependency	The logical requirement that the student's response to the test situation be dependent upon actually reading the test passages is met by the nature of the task, i.e., completing deletions. There are no questions apart from the passages.

Table 5.1 (Continued)

Property	Advantage
Automated generation of items and tests	The objective, generative nature of both the item and test format makes possible the automation of item construction and test assembly and printing. This makes for both speed and economy, which in turn will permit the use of more complex but more useful evaluation designs in the schools.
Flexible resource	The cloze passages are part of a flexible test development resource called the Test Development Notebook (TDN). Instead of providing a set of fixed tests, the TDN offers a collection of materials allowing rapid and economical assembly of large numbers of special purpose tests to fit a variety of evaluation needs. This format mitigates the problem of maintaining test security in large scale policy-oriented evaluation studies of reading and contributes to economy in test assembly.

Objective Item Format

The objectivity or reproducibility of test construction procedures has been a major criticism of norm-referenced measures of reading comprehension. Norm-referenced measures usually lack an explicit theory of comprehension and objectivity or reproducibility. The development of the multiple-choice cloze format in the TDN has involved a constant and, to date, largely successful effort to improve and maintain objectivity. Objectivity here, as noted, is important because, given other conditions, it enhances the possibility of repeatedly generating a test that incorporates an unbiased sampling of the content and behaviors of the universe of interest. The presumed unbiased nature of the test is, furthermore, traceable. Others interested in the operations defined by the test may generate similar or comparable tests, or the test may be generalized to other relevant behavioral domains in the course of extending or studying the underlying construct.

Effective solutions to the problem of objectivity in the generation of domain-referenced test items are offered in separate models by Hively et al. (1973) and Bormuth (1970). The generative item format represented by the multiple-choice cloze format used here is an application of Hively's concept of an item form in the domain-referenced testing model: (a) the multiple-choice cloze format constitutes the fixed or standard structure which contains one or more variable elements, and (b) the various unclozed passages and the distractor lists available for item and test construction are the replacement sets for those elements.

At latest study, the multiple-choice cloze format seems to offer the potential of being almost wholly objectively reproducible. Several modifications in procedure and format now under consideration will reduce potential biases in passage selection that may have resulted from earlier, unnecessarily rigid limitations on passage length; the use of titles on passages; and possibly insufficient unmutilated context at the beginnings and ends of passages.

The current rule-based procedure for conversion of passages to the multiple-choice cloze format seems to offer additional potential for computerization, thus further approaching the possibility of reducing the item form to an algorithm. This computerization would presumably string together separate programs for readability analysis, conversion of passages to the mutilated format, and the generation and assignment of distractors from the word lists.

As the objectivity of the multiple-choice cloze item construction and test assembly procedures of the TDN is further improved, some minor modifications of format will undoubtedly result. The current set of item analysis data is expected to contribute substantially to determining such modifications. Experience with assembling and using the test to date has also shown (see Chapter VIII) a number of areas where objectivity can now be enhanced.

Domain-Referenced Content

The passages for the multiple-choice cloze component of the TDN were systematically drawn from clearly defined and relevant domains of written discourse. The selection of relevant domains was aided by reference to the Hansen and Hesse effort in Madison, Wisconsin (1972) to build a domain-referenced test based on the standard cloze. In that effort parents and teachers identified relevant domains of written discourse on the basis of frequency of use or importance in the school and community for students in grades 4 through 12. The present effort improves on the relevance and specificity of these domains by extending the domains and levels to grade 1 and by incorporating readability as an additional defining characteristic.

The domain-referenced model is intended to support generalization from test scores to relevant domains of application. Theoretical and empirical clarification of the concept of literal comprehension, by contrast, is potentially indicative of defaults in the processes underlying literal comprehension. The ability to specify both the process of comprehension and the circumstances of its expression (i.e., the classes and levels of written discourse involved) in a test constitutes the basis for using comprehension test scores in decision-making in reading instruction.

Using the domain-referenced model as a basis for assembling the variety of passages for use in the TDN is a deliberate attempt to maximize the relationship between the test situation and program content. Program content includes relevant skills and materials involved in reading situations in the school and community. Maximizing this relationship should, in turn, enable someone to produce tests of reading comprehension that are maximally sensitive to some of the most important outcomes of reading instruction.

Such tests would be designed so that a given level or test form is suitable to the reading abilities of a given student population and so that the content of the test is relevant to and reflects the changing nature of the reading experiences of that group over time. A survey test for first graders, for example, would contain a range of passage difficulty that would reflect the range of written discourse relevant to first graders in the school and community. A set of tests assembled for first graders according to this principle would, at one point in time, theoretically generate the distribution of mean passage scores depicted in Figure 5.1.

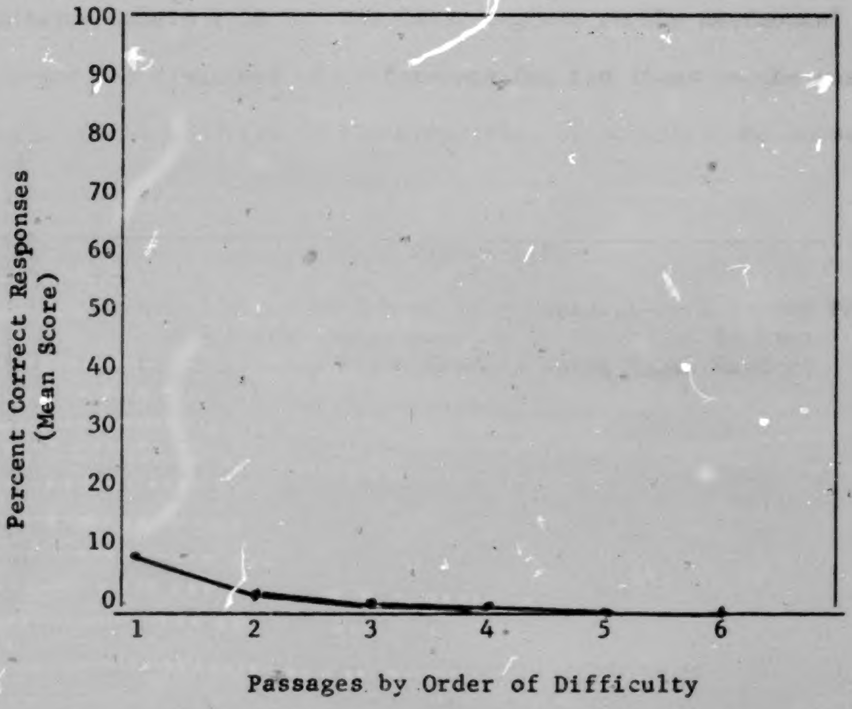


Figure 5.1. Ideal distribution of mean passage scores on survey test in reading given at beginning of grade 1.

Consider that the test contains six multiple-choice cloze passages ordered by difficulty or readability level. At the beginning of grade 1, as shown in Figure 5.1, the mean score on passage 1 is less than 10%, and it drops still lower on the other passages. This would be the expected performance of most first graders with relevant reading passages in a September test administration: most of them would not be able to apprehend the literal meaning of even the simple 25-word passages at readability level 1.

Figure 5.2 repeats the information of Table 5.1 and further demonstrates ideal or expected passage scores at later points in time for a test that accurately reflects experience with reading materials. According to this illustration, by the middle of grade 1 the mean passage score at readability level 2 is about 75%, whereas in September it was near zero. By the end of grade 1, the mean passage score at readability level 3 stands near 75%.

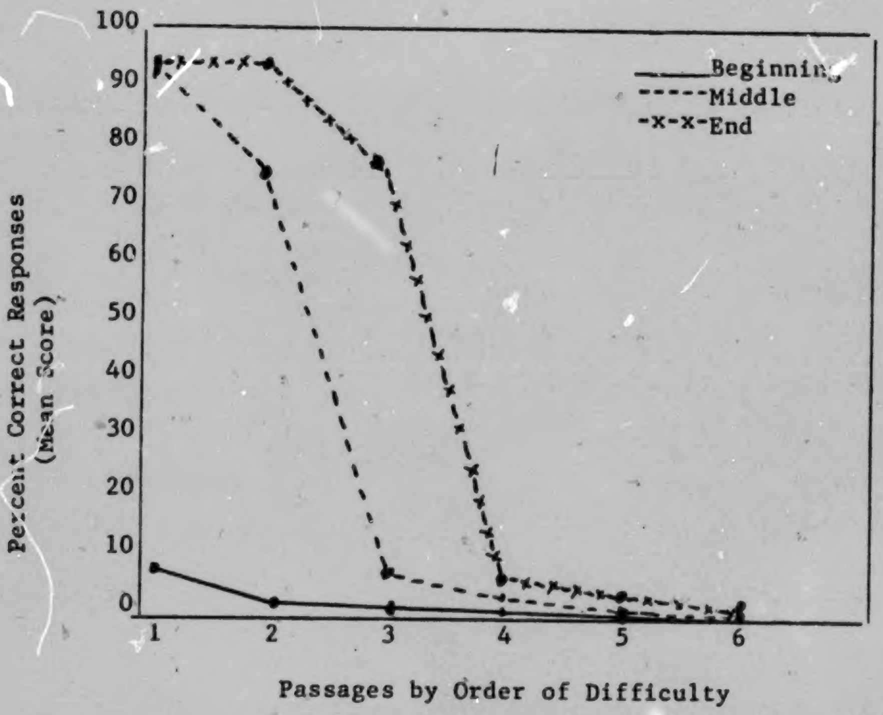


Figure 5.2. Ideal mean passage scores for survey test in reading given at beginning, middle, and end of grade 1.

An appropriately designed test of literal comprehension will thus generate a series of distributions over time which will reflect the increasing ability of a given student population to apprehend the literal meaning of increasingly difficult passages and more diverse domains. The nature of these performance distributions is a function of the time between test administrations, the degree of relevance of the content and the difficulty levels of the passages selected for the test, the ability of the students, and the degree of learning that occurs in the time framework of the test. The ability of such a test to measure certain broad effects of reading instruction and experience is theoretically maximized when the test is designed to generate the distribution shown in Figure 5.2 for each level of a given instructional system.

The ability to generalize from a score distribution like that shown in Figure 5.2 to one or more content domains is dependent on specifying the readability and content characteristics of both the test passages and the domains from which they were drawn. The description in Chapter IV of the design and assembly of the multiple-choice cloze component of the TDN showed that specification of the readability and content characteristics is well underway in the current development effort. Each of the cloze passages in the current set of testing materials has a Dale-Chall or Spache value identified by a two digit number, and considerable progress has been made in specifying the distribution of readability levels in each content domain in the TDN. What remains to be learned is the passage performance criteria to be applied to passages at a given level of reading development for the student population (e.g., Does 80 percent correct on a passage signify the passing criterion to be applied to a given population of students and passages?).[1]

---

[1] Some studies (e.g., Bormuth, 1971) suggest that the efficiency level of readers improves with age and experience, thus suggesting the development of age-graded performance criteria.

Unidimensionality

The multiple-choice cloze constitutes a procedure for measuring a type or level of reading comprehension that has its basis in cognitive theory and in generative theories of language (Smith, 1975). The test or item format is an attempt to measure comprehension at one level of cognition, i.e., the literal meaning of a written message. This is distinct from such other types of comprehension as induction, deduction, or evaluation, where the reader is required to go beyond the literal meaning.

The design of the test format is basically coherent with the ongoing act of comprehension, which is characteristically rapid and focused on the processing of relatively large informational units. The structure of the passages in a given test is unchanged by the multiple-choice cloze format. The act of responding to the word choices remains focused on the meaning of the passage, interrupted only by periodic deletions. In this act, the reader predicts the words that correctly replace these deletions to complete the intended meaning by drawing upon his own relevant cognitive structures, unique or otherwise, and not, as may be the case with other formats, those in the mind of a test-maker.

The behavior measured by the multiple-choice cloze format, termed literal comprehension, is one of a small number of hypothetical factors which account for how a reader might process different types of written materials under different circumstances. Literal comprehension is probably that component of comprehension that is most heavily affected by instruction and experience during and beyond the period normally given to the formal reading program. The focus on a measure of literal comprehension, therefore, promises to result in a more sensitive methodology for testing the actual or achievable outcomes of reading instruction across much of the term of public education.

Furthermore, the organization of the present work affords a unique opportunity for studying the meaning and development of literal comprehension in a response context that is essentially constant across age and grade. The multiple-choice cloze format offers basically the same stimulus to different age-graded respondents, the only differences lying in the complexity or difficulty of the passages used at different levels. The Rasch measurement model, moreover, is currently being experimentally applied to cross-sectional data on the test to provide assurance of unidimensionality. This may be one of the first reading tests where dimensionality was studied at the level of the item format in a developmental context. Certainly, existing tests of comprehension do not provide the assurance that apparently similar types of multiple-choice questions measure the same factors across test levels.

Equal-Interval Scaling

The application of the Rasch measurement model to the total pool of clozed passages, using the technology provided by Wright and Mead,[2] will result in the calibration of all passages on a single scale with equal-interval properties. Because of the way it was assembled, it will be possible to fix this scale in relation to meaningful upper and lower limits that will give the scale the properties needed for cost benefit analyses and other important evaluation purposes. The lowest level on the scale, for example, will be fixed just above the easiest of passages available in pre-primers, while the upper level of the scale will be fixed to the more difficult passages sampled in twelfth grade and in selected adult reading materials. The total scale of passage calibrations will thus cut across the total range of passage readability that is relevant to the grade 1-12 student population.

---

[2] Personal communication.

1-12 student population.

The equal-interval scaling of the test passages provides the possibility of determing how much of the total reading curriculum in terms of literal comprehension has been achieved at a particular point in time. For example, if the readability levels of basal readers and related texts range across some 20 readability levels, and a given group has attained a test score that represents mastery at the uppermost of these levels, then it may be concluded that the level of achievement with this domain is 100%. If it was achieved in 9 instead of 10 years, then the level of efficiency might be given as 110%.

The foregoing scaling properties allow for the potential development of a variety of meaningful scores that will appropriately transform the base score of any test assembled from the item pool by taking into account the amount of instructional time and the amount of content achieved. Content is defined as the number of domains at specified readability levels (e.g., presumably by 6th grade, which is the half-point of school time, some specifiable proportion of each relevant domain of written discourse should be achieved).[3] The fineness of the calibrations of passages in the TDN which application of the Rasch model will achieve, moreover, will make the item pool appropriate for the assembly of tests with relatively fine or coarse calibrations as required for different assessment purposes. The original item pool was assembled using equivalent passages that were calibrated (by the use of readability formulas) by half-grade readability intervals. The Rasch calibration should yield a much finer scale of passage calibrations.

---

[3] The sampling of readability levels of basal readers and literature texts rarely resulted in materials above level 20, while materials in the content areas were generally shown to be more difficult at the upper grade levels. One may conceive of a derived score which would express some average of the proportions of each domain that should be read by a given individual, based on the distributions of readability levels by content domain and grade or age.

The Dale-Chall scores of the passages at given points on the projected Rasch scale will provide a partial basis for generalizing a test score to a given universe of written discourse. The additional referents required are, as noted, the mastery performance criteria to be applied to a given level of the population and distributions of readability scores for the domains of written discourse relevant to a particular level of the student population.

Passage Dependency

As noted in Chapter I, the validity of several well-known tests of reading comprehension has been seriously challenged on the issue of passage dependency--the tendency for students to obtain scores well above chance without reading the test passages. Some authors have proposed that this issue be handled by redefining comprehension as information gain (Bormuth, 1970) or as a residual gain (Rankin and Dale, 1969) rather than by making better tests. These procedures attempt to remove from the test score the influences of specific and general knowledge and various other test-taking strategies that operate on test questions independent of the test passages.

Though one might wish to define new learning in this way, i.e., as information gain, redefining comprehension in such terms seems to create new problems. One has only to consider for a moment the virtual impossibility of distinguishing between that part of the test score that represents prior knowledge of the reader and that part which represents new learning. How does one ask a question about a passage in the pretest situation which does not involve the interplay of old and new information or knowledge?

It does not seem desirable, necessary, or possible to attempt to remove the influences of specific, general, or idiosyncratic knowledge from comprehension test scores. According to the position taken here, all meaning exists in the reader and reading or comprehending necessarily involves bringing such meaning to bear on comprehending a passage. The best available solution to the problem of passage dependency, therefore, involves eliminating as much as possible the influences of additional meaning or other irrelevant informational cues offered by the test situation. Such influences are necessarily present in the typical multiple-choice questions in tests of comprehension (e.g., idiosyncratic meanings are introduced by the ways in which the test-writer interprets a given passage), but it would be difficult or impossible to eliminate them without also seriously affecting the measurement of comprehension.

The multiple-choice cloze format appears to avoid the issue of introducing into the test situation idiosyncratic meaning that both affects the test score and interferes with the student's attempt to process independently the information in the test. Theoretically, this was accomplished by eliminating semantically plausible distractors from the word choices given for each deletion and by voiding syntax as a basis for choosing the correct response. The effect of other cues (e.g., distractor length) on the test score is another issue that will be handled by the distractor review process.[4] However, only empirical study will determine the overall degree of success obtained in handling this problem.

Automated Generation of Items and Tests

As noted, the cloze component of the TDN is not a fixed test but will be a bank or collection of calibrated passages intended for various evaluation purposes. Also, the flexible notebook format is an effective device patterned in principle after the item banking experiments that have proven so successful in supporting the economical assembly and maintenance of tests for CAM in the Sequoia and Hopkins projects in California and Minnesota, and in various installations in New York State (cf. Gorth et al., 1975).

---

[4] The formation of part-of-speech word lists in the distractor generation process, for example, is one of several strategies designed to eliminate the possibility of selecting the correct word without reading the passage.

Experience in these projects shows that approximately half the cost of providing achievement scores to students is in the development, design, assembly, maintenance, and production of tests (Gorth et al., 1975). In addition, the more sophisticated and useful evaluation designs are not even economically feasible unless some way is found for systematizing the generation, maintenance, and production of the required tests. For example, it is obvious that state and district level evaluation models can be greatly improved by eliminating the tenuously secure standardized achievement batteries now used in favor of the multi-matrix sampling approaches which incorporate a large array of test forms and also produce data that are more broadly representative of a system's goals. However, the lack of effective technical support for mounting such designs has most likely been the major factor preventing their implementation at state and local levels.

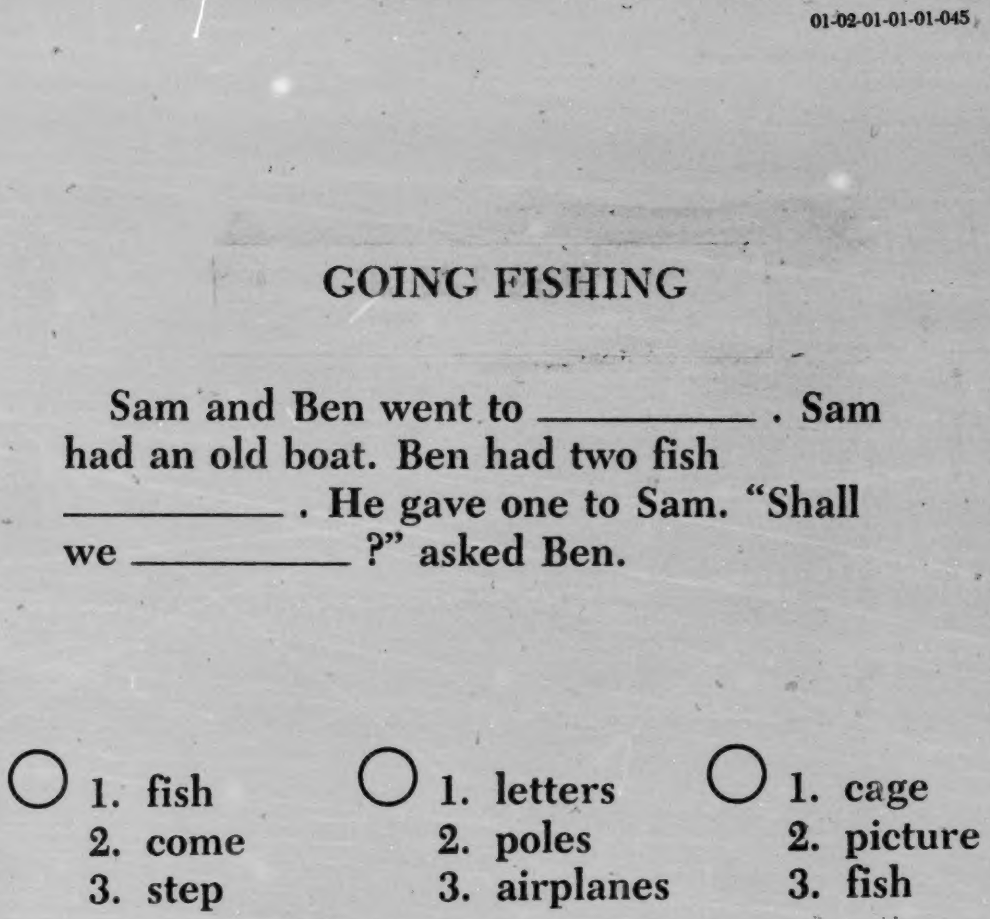
The design of the cloze and other components in the TDN anticipated the effective use of technology to support the development and maintenance of a given bank of items and the assembly and economical production of large numbers of finished test forms. Because of the form taken by the cloze component of the TDN, it has also become feasible to partially automate the item generation procedure. The discussion on objectivity or reproducibility indicated that the processes of producing a clozed passage might be computer programmed once the passage was selected.

The technical support for storing, reviewing, assembling, and printing multiple copies of passages and items in the cloze component of the TDN is based on the use of the Merganthaler V-I-P (Variable Input Phototypesetter) Model 7245-3. This phototypesetter reads a paper-punched tape which utilizes the standard TTS (teletypesetting) 6-level code. The TTS 6-level code enables the selection of 96 characters (alphanumerics) and 22 command codes specifying typesetting parameters (such as fixed and variable spacing) and machine control functions (such as shifts and line endings). The Merganthaler "reads" the command codes and then exposes the selected characters in the command format onto photo-mechanical paper. From this, a printing plate is produced for the rapid duplication of multiple copies.

The process of converting the cloze format passages and items to this paper-punched tape medium and inserting the programming instructions on layout for printing is currently underway. This process interfaces directly with printing (eliminating conventional procedures for typing drafts), and also supports a system of easy storage and editing. The 1,374 cloze passages and items are being stored on approximately 350 tapes. After evaluation and field testing, individual items can be edited, and passages altered or replaced through a video-correction terminal in the State Education Department.

This procedure also provides an economical means of generating various test forms. Depending on field-testing, and eventually user needs, test forms can be generated through the selection of the appropriate tapes from the bank.

Figure 5.3 is a copy of one multiple-choice cloze passage produced on the Merganthaler phototypesetter. The copy is in Caledonia-Bold



01-02-01-01-01-045

GOING FISHING

Sam and Ben went to __________ . Sam had an old boat. Ben had two fish __________ . He gave one to Sam. "Shall we __________ ?" asked Ben.

○ 1. fish
2. come
3. step

○ 1. letters
2. poles
3. airplanes

○ 1. cage
2. picture
3. fish

Figure 5.3. A copy of a cloze passage as produced by a phototypesetting machine

typeface, 14-point size. The mark-up sheet for this passage is shown in Figure 5.4.

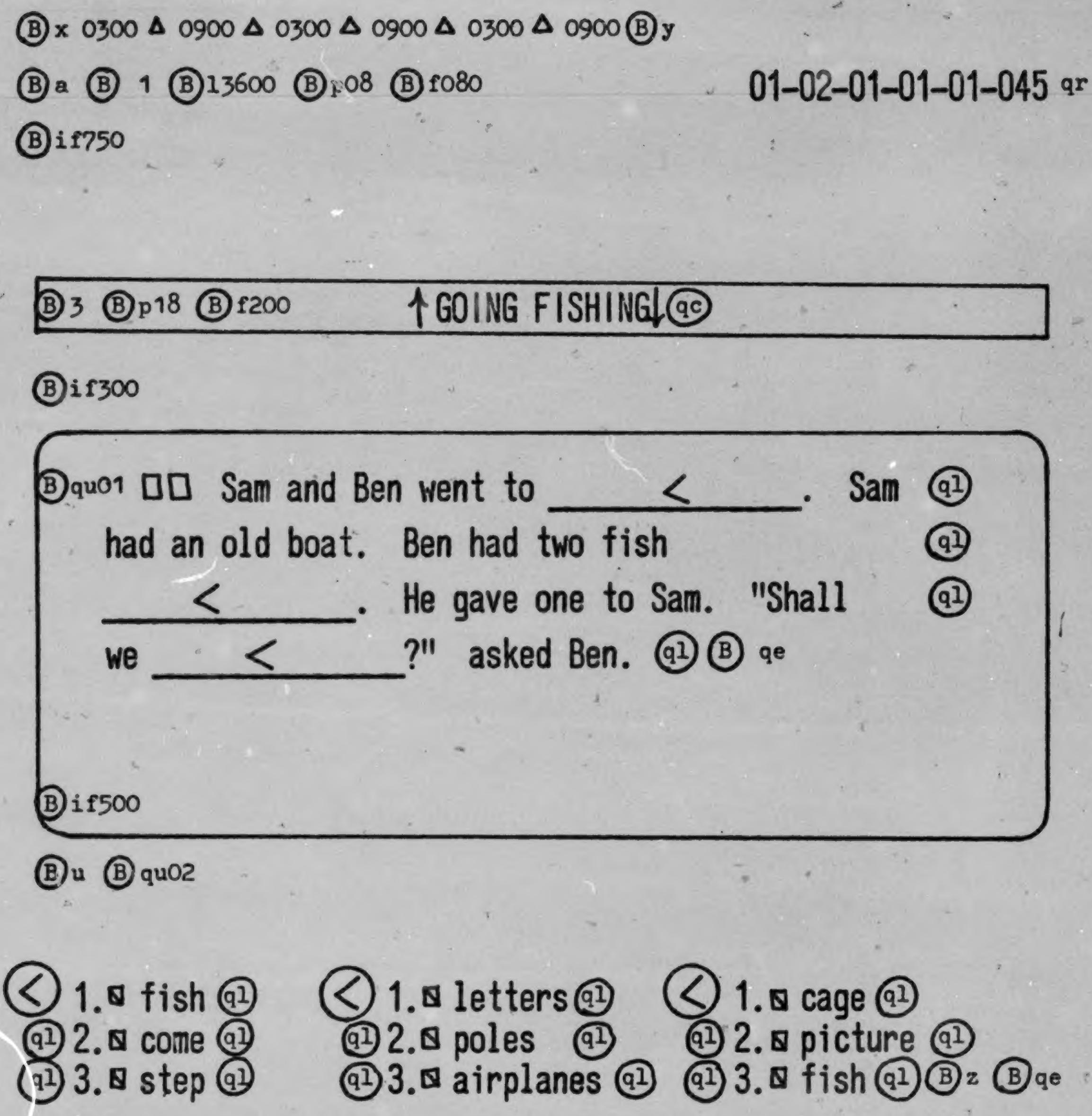
The foregoing process is the core of a test assembly procedure that is now currently operational and effective for the problem it addresses. It will ultimately be integrated with the computer to further improve speed and economy and also to interface test production with the process of analyzing response data.[5] The computer will store information on all passage characteristics on a disc or tape file and will provide a program for the selection of passages on the basis of several simultaneous criteria. Presumably, these criteria will include the range of calibrations desired in the test, the number of passages, the content areas to be sampled, and the grade level(s) of the student population.

Once the content of the test(s) has been specified, the program will be capable of generating data decks that identify the characteristics of the test and determine how it is to be scored. A test generated by such a system will be provided in the required number of copies and will be scored and processed for reporting purposes. It is expected that the development work on the cloze component of the TDN will be carried to this point.[6]

---

[5] Currently, the tapes can be conveniently filed and used for the assembly of large numbers of tests without the aid of a computer. The entire set of 350 tapes is being reviewed and edited prior to the production of 1,000 copies of the item bank. These are to be used in local assembly of test forms by constructing test form masters directly from hard copy.

[6] Other files, such as the student files, will need to be set up before the test is scored and a report produced. However, the anticipated production of the item scoring file, along with the test, will contribute to the speed and economy with which a given evaluation design can be mounted based on the production of a set of unique tests.



Ⓑx 0300 △ 0900 △ 0300 △ 0900 △ 0300 △ 0900 Ⓑy

Ⓑa Ⓑ 1 Ⓑ13600 Ⓑp08 Ⓑf080 01-02-01-01-01-045 qr

Ⓑif750

Ⓑ3 Ⓑp18 Ⓑf200 ↑GOING FISHING↓(qc)

Ⓑif300

Ⓑqu01 □□ Sam and Ben went to ____<____. Sam (q1)
had an old boat. Ben had two fish (q1)
____<____. He gave one to Sam. "Shall (q1)
we ____<____?" asked Ben. (q1) Ⓑ qe

Ⓑif500

Ⓑu Ⓑqu02

(<) 1.▨ fish (q1) (<) 1.▨ letters (q1) (<) 1.▨ cage (q1)
(q1) 2.▨ come (q1) (q1) 2.▨ poles (q1) (q1) 2.▨ picture (q1)
(q1) 3.▨ step (q1) (q1) 3.▨ airplanes (q1) (q1) 3.▨ fish (q1) Ⓑz Ⓑqe

Figure 5.4. Example of a cloze passage prepared for keypunching on a teletypesetting machine showing textual copy and typesetting commands.

Flexible Resource

Perhaps the principal advantage of the multiple-choice cloze materials, in the context of the TDN, is flexibility of application. The anticipated range of application of the cloze component of the TDN is defined in terms of three levels of evaluation identified in column 1 in Table 5.2. They are survey testing, achievement monitoring, and diagnostic or tailored testing. The key decision-makers at each level of evaluation are given in column 2. The time frame of test administration in a level of evaluation is shown in the third column. The fourth column gives some brief examples of the purpose of the test administration, and the final column shows examples of the types of decisions that each group might make, given the kinds of data that result from a type of testing. In practice, no one level of test information is used exclusively by any one decision-making group. Rather, information from testing becomes progressively less useful as it is more removed from its intended primary reference group.

The level of testing that is undoubtedly most familiar to most educational decision-making groups, professional and client alike, is the survey test usually associated with the widespread annual administration of standardized achievement tests. The example of survey testing given later in this chapter carries the same intent: to assess the status and development of the student population in terms of major educational outcomes and domains of application--in this case, literal comprehension as applied to specified categories of written materials.

Survey testing. The survey-testing design using the cloze passages presents each student with a sample of passages, such that a broad range of relevant written discourse is tested against specified populations in a school or district. The data resulting from such a design will provide estimates of

Table 5.2

Range of Application of the Multiple-Choice Cloze Materials

Level of evaluation	Key decision-makers	Time frame	Purpose	Decisions
Survey testing	Administrators	Annual or bi-annual	Assess comprehension in a range of levels and domains across student population.	Allocate resources; determine effectiveness of reading program(s) (by district, buildings, levels, or other units) over a long-term period.
Achievement monitoring	Principals, teachers and students	Periodically (e.g., every 5-10 weeks)	Assess growth of comprehension within a level and domain.	Allocate instructional time and effort continuously throughout a course; determine student progress; select materials for a course or student; assign students to a level in the system.
Diagnostic or tailored testing	Teachers, students	As needed	Assess comprehension level at one time on a skill-by-skill basis.	Determine a student's level of reading; find materials suited to a student's reading level.

the status of literal comprehension in relevant domains of written discourse by grade level, building, or attendance area. For example, the results could show that, across a sampling of 25 different basal reader systems, 20 percent of first graders scored 90 percent or better on the highest level of passage difficulty in June. At higher levels in the educational system, the same type of performance estimates could be shown for a broader set of relevant materials. For example, the survey test for middle-school students would likely sample across reading, language arts, science, social studies, reference materials, consumer materials, and so on. Such a design might thus make use of the principles of multi-matrix sampling by sampling the various item domains available to obtain the broadest possible representation of content on the test.

Survey testing is not necessarily useful at the individual level, particularly when matrix sampling is involved, since any one student may receive a test composed of only a narrow sample of passage content and reading levels. Survey data are primarily used to generate group performance estimates aggregated to a particular level of interest, e.g., all fourth graders in a given building. Survey data based on the TDN multiple-choice cloze materials will thus enable administrators and program managers primarily to examine and follow the development of the total reading program from year to year. The associated decision-making will typically be broad and long-term. The district administrators will generally use the test results to identify needs in terms of student groups and problems with areas of written discourse. They will use the data to examine the development of literal comprehension over several years. They may ultimately begin to adjust the body of textual and other written materials to fit the reading needs of the school population. All of these and other decisions will be largely based on a new type of norm made possible by the structure of the testing materials. This is a norm that is referenced to a given category and level of written discourse that is exemplified by the statement: *75 percent of seventh graders achieved literal comprehension scores of 90 percent or better on a sampling of editorials from major newspapers.*

Achievement monitoring. The next level of evaluation referred to in Table 5.2 is achievement monitoring. Achievement monitoring is a newcomer to the practical context of classroom evaluation. Developed as a standard design in the project by Gorth et al. (1975), the basic elements of achievement monitoring are a set of parallel test forms and a longitudinal test schedule in which the tests are repeatedly administered without duplication to each student in a program. The forms are randomly administered at fixed intervals; e.g., with five forms and a bi-monthly test schedule, a given student might receive tests in the order: 1, 5, 3, 4, 2.

If the design of each test form included cloze passages with equivalent ranges of readability levels selected from the TDN, each test administration would yield an estimate of a student's level of literal comprehension based on the same standard. For example, each of five test forms could sample readability levels 1-5 for a class of first graders. The resultant data at each interval would yield individual and group performance estimates at each readability level; e.g., *in September 90 percent of the class achieved 90 percent or better at level 1, 80 percent or better at level 2 . . . and 20 percent or better at level 5.* The teacher using such data would be periodically looking for expected increases in literal comprehension at higher levels of readability as the course of reading instruction unfolded.

Diagnostic or tailored testing. The third level of testing referred to in Table 5.2 would use passages selected from the cloze component of the TDN to generate a test tailored for one-time or repeated testing of a single student. The cloze materials are ideally suited to the rapid assembly of such tests for the purpose of determining the level of the materials a student is able to comprehend literally, e.g., texts in the various content areas, newspapers, news magazines, consumer materials, etc. This testing would be useful whenever a new student entered the school and his level of reading ability was unknown. The resultant test scores from the cloze would indicate the levels of reading materials in each content area that were appropriate to the student's ability in the instructional and independent reading contexts.

Application of the Cloze Testing Materials in Practice

Thus far only one of the major evaluation purposes of the cloze materials--the survey--has been explored. An experimental application of the survey was administered in late May and early June 1975, to 5,000 grade 1-9 students in a school district in upstate New York. The remainder of this chapter presents the basic principles for the development of a survey design and a detailed description of the development and implementation of this first survey design. The design of this survey test was developed prior to any calibration of the cloze passages. The purpose of this survey test administration was to collect item and passage data which would provide a basis for a validity study, determine the adequacy of the cloze format, and investigate the utility of readability formulas for test assembly.

Design Principles

In "Sampling Plans for Domain-Referenced Tests", Millman (1975) presents the basic principles involved in assembling domain-referenced tests for varying evaluation purposes. Figure 5.5, adapted from Millman's article, illustrates these principles. Each cell in the figure represents the con-

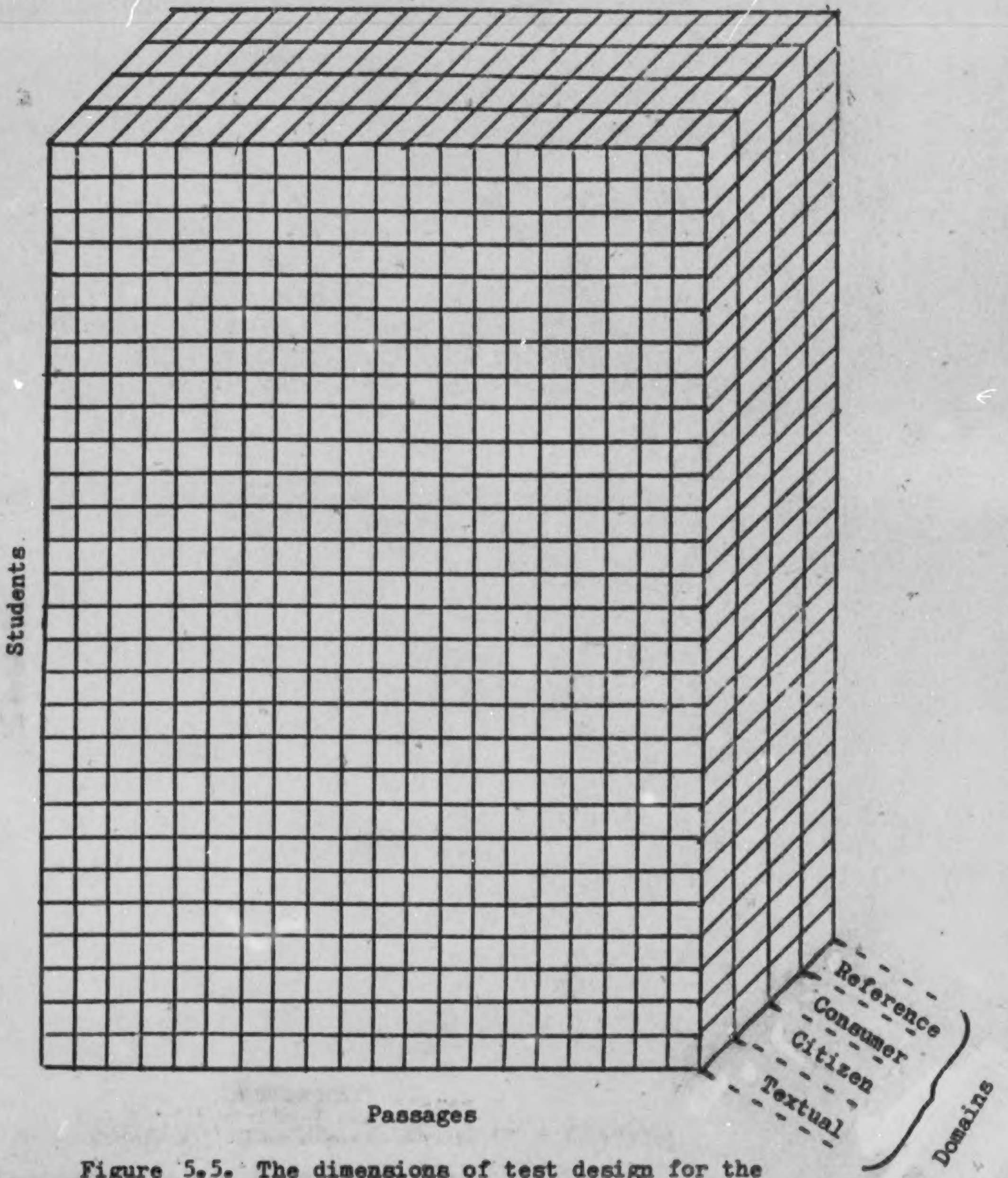

Students
Passages
Reference
Consumer
Citizen
Textual
Domains

Figure 5.5. The dimensions of test design for the cloze segment of the TDN.

vergence of a particular student, a passage or set of items, and a domain of written discourse. Tests required for given evaluation purposes are assembled through a sampling plan which operates on all three of these dimensions in the framework of time. The evaluator's problem is to select students and passages in domains relevant to the specific information needed.

For example, say a reading teacher needed a test to determine the appropriate level of reading text for a new sixth-grade student. Here the student dimension would involve only one student; one category of textual materials (i.e., reading/literature--not the entire textual domain) would represent the content sampling unit; and the passage domain would be a set drawn from the 353 available passages in this subcategory in the TDN.

The teacher might begin by randomly drawing a passage from each of the readability strata represented by difficulty levels 8-12 shown in Table 5.3. These passages would be used to compose a 50-item test which would probably encompass the student's reading level. The student's score on this test would be an estimate of his reading ability across all readability levels in the reading/literature subcategory. Assurance that this score is an accurate estimate of the student's ability to comprehend written materials in this category could be increased by drawing additional passages from the matrix in Table 5.3. Such passages would represent a narrower range of readability than the passages on the initial test, and the process could be repeated several times before the passages in a given cell would be exhausted.[7]

---

[7] At present, only readability information based on Spache and Dale-Chall formulas is available as a guide for passage selection. Eventually, passage selection will be based on the Rasch calibration of the total passage pool. The scale thus calibrated will still be referenced to readability scores for the sake of providing a technical basis for selecting texts and other written materials for a range of comprehension scores.

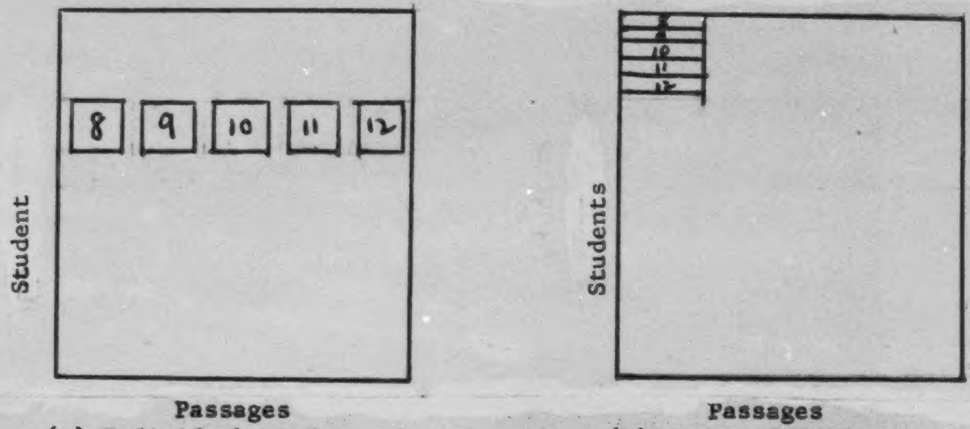
Every scale score or range of scale scores on the total test will be associated with a set of "equivalent" passages accompanied by readability score means and standard errors. These data are the primary bases for estimating the difficulty levels of textual or other written materials which a given student or group can comprehend with a certain degree of confidence.

Table 5.3

Multiple-Choice Cloze Passages in the Reading/Literature Textual Domain

Grade level	Readability level	Grade level of passage source										Totals
		1	2	3	4	5	6	7	8	9	10	
1	1	11										11
1	2	30										30
2	3	7	14									21
2	4		26									26
3	5			12	1							13
3	6			12	4							16
4	7			6	13	7						26
4	8				9	4						13
5	9				11	11	6					28
5	10				4	6	4					14
6	11					6	11	2				19
6	12					2	5	3				10
7	13						6	11	2			19
7	14						1	5	1			7
8	15							5	10	4		19
8	16							4	8	4		16
9	17								6	5	5	16
9	18								3	9	7	19
10	19									4	4	8
10	20									4	8	12
11	21										3	3
11	22										7	7
12	23											
12	24											
13	25											
13	26											
14	27											
14	28											
Totals:		48	40	30	42	36	33	30	30	30	34	353

The foregoing sampling design for assembling a cloze test based on passages from the TDN for an individual student is illustrated in Figure 5.6(a). This diagram depicts one student being administered a stratified random sample of passages from one content domain or category in the TDN. The strata, indicated by a row of small boxes, are the readability levels shown previously in Table 5.3. Figure 5.6(b) illustrates nearly the same plan as 5.6(a), but here a group of students randomly drawn from the unit of interest (e.g., classroom, grade-level, etc.) is given the same test form. Readability strata are now indicated by horizontal lines in the small box. Figures 5.6(a) and (b) represent the simplest sampling designs that might be drawn for the cloze segment of the TDN.



(a) Individual student

(b) Group of students

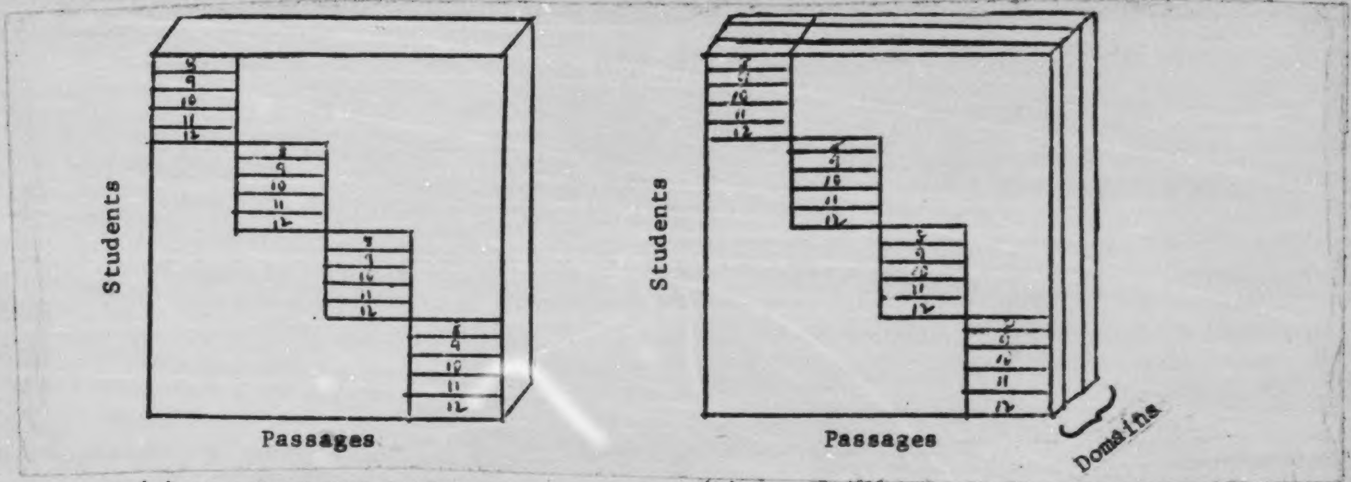
Figure 5.6. Simple sampling designs for the cloze segment of the TDN.

The more complex matrix sampling designs that are possible with the cloze component of the TDN are illustrated in Figures 5.7(a) and (b). Figure 5.7(a) generalizes the one-test-form/one-group plan in Figure 5.6(b) to a matrix sampling design in which non-overlapping random samples of passages are assembled into parallel test forms, and each test form is administered to a different, randomly constituted group of students. On both sampling dimensions, random stratified sampling is used. Each parallel test form samples from the same readability strata, and test forms are assigned randomly to students within ability strata in each experimental unit of interest--classroom, grade level, and so on.

The sampling design presented in Figure 5.7(a) is the standard CAM design. It is technically defined as a longitudinal, multi-matrix sampling design. Application of the standard CAM design would involve assembly of several sets of parallel test forms to cover the range of ability in the groups of interest. For example, evaluation of growth in comprehension in sixth-grade classrooms might require three to five sets of test forms, each set encompassing a restricted range of readability (e.g., set one might have a range of 6-10; set two might have a range of 11-15, etc.). A student assigned to a set of such test forms would receive them at fixed intervals throughout the reading program. The differences in domain scores at each data point would provide a basis for estimating growth or development in literal comprehension.

Design 5.7(b) illustrates the sampling scheme for survey testing. This design may be viewed as a one-time application of the standard CAM design for each content domain of interest. A survey testing system might thus include a series of parallel test forms for all of the four dimensions or nine content categories in the TDN, with each set of test forms administered to a different randomly constituted population of students. The design could be further varied to yield different levels of information on a population, depending on the content domain or category. For example, each individual in the unit of interest (e.g., a grade level in a district) might receive one of several test forms in the first category (reading/literature) in the textual domain. This would yield an estimate of each individual's domain score for that category. Thereafter, in a second testing session, each individual might receive one additional test form in one of the remaining domains or categories. In this way, a survey design could be mounted to yield both individual and group data as required.

There are a number of complex considerations involved in developing sampling plans for a survey based on the TDN. A plan actually consists of several sampling designs, each developed for a given level of the population (e.g., Level I, grades 1-3; Level II, grades 4-6; Level III, grades 7-9). The passage readability levels of each survey level should overlap. There may be a relatively large number of test forms, depending on the number of content areas surveyed, and the test administration schedule will probably require a computer to effectively accomplish the assignment of test forms to populations.



(a) Standard CAM design (b) Survey testing design

Figure 5.7. Multi-matrix sampling designs for the cloze segment of the TDN.

An Applied Survey Design

The experimental survey design based on the multiple-choice cloze segment of the TDN applied the model illustrated in Figure 5.7 (a) to the reading/literature category in the textual domain. The initial purpose of this effort was to obtain item analysis data on a large segment of a major content area in the TDN. Subsequently, the Rasch analysis program supplied by Wright and Mead (1975) was experimentally applied in the analysis of items and passages. The Rasch data was also used (in conjunction with Wright and Mead) to explore the unidimensionality of the characteristic measured by the multiple-choice cloze.[8] It was expected that the results of these analyses would be used as a basis for making any needed adjustments in the cloze format prior to starting the major effort on calibration and validation.

Because plans for validating the multiple-choice cloze materials included concurrent use of modified wh-items (see procedure for producing wh-items, Appendix A), the survey design included a substantial number (approximately 1,000) of wh-items. (Note: The test in the multiple-choice cloze format was called Cloze Exercises, and the test in the wh-item format was called Literal Comprehension, Details.) The design incorporated the largest number of test forms and items or passages that could be administered to the survey sample and yield stable data on each item or passage.

The test administration sample is defined in Table 5.4. For each grade from grade one through nine between 500 and 750 students were tested. Each test was designed to be administered in a 40-minute class period. Since class period time restrictions were flexible in grades 1 through 8, students in these grades had ample time to complete the tests. Class-period schedules were not flexible in grade 9; thus, as Chapter VIII will note, some ninth-grade students did not have sufficient time to complete the Cloze Exercises. Each student had some advance training in the cloze format. This ranged from several 20-minute training sessions in the primary

---

[8] Dr. Steven Kidder, a BSCR staff member, has studied the Rasch model data resulting from this test administration of the multiple-choice cloze with Dr. Benjamin Wright and his graduate assistant, Ronald Mead, at the University of Chicago.

Table 5.4

Number of Students Tested per Form, Cloze Exercises and Literal Comprehension, Details Test

Cloze Exercises (N = 5,264)[a]						Literal Comprehension, Details Test (N = 5,197)[a]					
Level I		Level II		Level III		Level I		Level II		Level III	
Form	N	Form	N	Form	N	Form	N	Form	N	Form	N
1	128	13	147	25	167	37	127	49	147	61	163
2	126	14	152	26	164	38	124	50	152	62	162
3	130	15	153	27	160	39	126	51	148	63	164
4	124	16	152	28	161	40	121	52	152	64	161
5	126	17	146	29	158	41	119	53	145	65	165
6	126	18	151	30	165	42	122	54	144	66	166
7	127	19	152	31	158	43	124	55	145	67	154
8	127	20	148	32	163	44	124	56	149	68	163
9	129	21	152	33	166	45	131	57	147	69	164
10	127	22	152	34	159	46	123	58	148	70	156
11	120	23	148	35	163	47	121	59	157	71	163
12	123	24	149	36	165	48	121	60	145	72	154
All	1,513	All	1,802	All	1,949	All	1,483	All	1,779	All	1,935

Note. Total N = 5,722. Distribution by grade: 1, 551; 2, 536; 3, 524; 4, 580; 5, 649; 6, 692; 7, 751; 8, 709; 9, 730.

[a] Not all students were present for all tests.

grades to a 15-minute session in grade 4 and above.

Sampling Design for the Cloze Exercises

The Cloze Exercises were composed of three levels of parallel test forms: Level I, for grades 1-3; Level II, for grades 4-6; and Level III, for grades 7-9. The first problem in designing the survey was determining how many readability levels to include in each testing level so that the test would encompass the lowest and highest achievement levels of most of the intended population. Table 5.5 shows how this problem was resolved. The first testing level, Level I, was assigned the first 10 of the 28 readability levels covered by the cloze passages in the TDN. Passages in this range begin with simple 25-word excerpts from grade 1 basal readers and extend to 70-word passages, sampled from grade 4, 5, and 6 readers. Level II was assigned readability levels 5 through 16; passages in this range come from materials for grades 3 through 9. Level III was assigned readability levels 11 through 22, with passages coming from materials for grades 5 through 10. (Note: All passages at a given readability level were considered equivalent regardless of the grade levels of their sources.)

With these readability ranges established for each testing level, passages were then sampled from consecutive pairs of readability levels, with the exception that readability levels 1 and 2 were each considered separate sampling units. The sampling design for all three testing levels is shown in Table 5.5. A set of 12 parallel test forms was assembled for each level (i.e., Levels I, II, and III) by randomly sampling passages repeatedly and without replacement from each sampling unit assigned to that testing level. Table 5.5 shows the number of passages drawn for any test form from a given sampling unit, i.e., readability level or pair of readability levels. Thus, a test form for Level I sampled passages from readability levels 1, 2, 3 and 4, 5 and 6, 7 and 8, and 9 and 10.

Table 5.5

Sampling Design for Survey Test in the Textual Domain--
Reading/Literature for Cloze Exercises

Sampling unit: Readability level	Passage Pool by Grade Level of Source										Passages per unit
	1	2	3	4	5	6	7	8	9	10	
1	11										1
2	30		Level I								1
3	7	14									1
4		26									
5			12	1							1
6			12	4	Level II						
7			6	13	7						1
8				9	4						
9				11	11	6					1
10				4	6	4					
11					6	11	2				1
12					2	5	3	Level III			
13						6	11	2			1
14						1	5	1			
15							5	10	4		1
16							4	8	4		
17								6	5	5	1
18								3	9	7	
19									4	4	1
20									4	8	
21										3	1
22										7	

In the test administration, the test forms were systematically distributed to obtain equal numbers of respondents on each test form by classroom. This was done by packaging the forms for each classroom in numerical sequence, repeating the order until the required number of tests had been packaged for a classroom, and starting the sequence for the next classroom where the previous one had left off.

Sampling Design for Literal Comprehension, Details Test

The Literal Comprehension, Details test used in the survey was assembled using the design shown previously in Table 5.5 for the selection and assignment of passages to test forms. The one variation was at Level III where, since available materials only encompassed readability levels 1-20, readability levels 11 and 12 were regarded as separate sampling units.

Once the passages had been identified, five of the multiple-choice wh-items accompanying them were selected by a process which ensured equal representation of item types across test forms and readability levels. Equal representation was not possible at readability levels 1 and 2, where there were few adverbial items (i.e., how, when, where). This procedure yielded three sets of twelve 30-item test forms, Level I, Level II, and Level III. Except for the modification noted at testing level III, a Literal Comprehension, Details Test form in a given testing level included the same number of passages and range of readability levels as a test form at the same testing level in the Cloze Exercises.

To verify the high level of passage dependency expected with the wh-format, a "Part II" section was added to the test. Each Part II section consisted of 12 wh-items without their related passages, two for each readability level on Part I of the test. A table of random numbers was used to assign items to test forms, with the conditions that the items on

Part II of any given test form should be unrelated to passages on Part I and that each wh-item type should be represented at least once on a test, but no more than twice.

Other Data in the Survey Design

Once the resultant data from the Cloze Exercises and the Literal Comprehension, Details test were collected, arrangements were made to obtain the response data for the same students on selected subtests of the California Achievement Test (CAT) given in grades 1-8 in May, 1975. This data set included all of the item responses for the reading and language arts subtests and an IQ score derived from sections of the achievement battery. These data permitted expansion of the research perspective to include selected validity studies involving the Cloze Exercises and the Literal Comprehension, Details test.

Conclusion

Previous chapters have presented a rationale for the multiple-choice cloze format as a measure of literal comprehension and have further described how this format was applied in the development of a testing system, referred to as the TDN, which is ultimately intended to supersede conventional comprehension testing systems with the more flexible and potentially more useful approach implied in the domain-referenced testing model. An operational, domain-referenced testing model is not based on a set of inflexible, fixed tests. Rather, it has the facility to generate a test for virtually any evaluation purpose by the working through of an algorithm which can produce any number of test items as needed to survey a domain. Sampling procedures are then applied to these item domains in an evaluation design that attempts to eliminate, as far as possible, various sources of bias. Particularly evident in the domain-referenced model is an improved potential for eliminating much of the content or item bias that has been noted as a serious problem in standardized tests of reading comprehension.

Unfortunately for the development of this project, there are few operational models that can serve as detailed guidelines for the design of domain-referenced tests, particularly for the development of testing systems that are referenced to very large stimulus and response domains, such as the virtually infinite field of comprehension as applied to the domain of written discourse. This project is an attempt to apply the domain-referenced testing model to very limited but still very large segments of that domain. In the course of achieving this goal, the theoretical elements of the domain-referenced model were carefully followed. Insofar as possible at this time, the psychological meaning of the response required by the multiple-choice cloze item format has been elaborated.[9] This item format has been brought to a highly objective state where it now seems to have the essential objective, generative characteristic required by the domain-referenced model. Further developments along these lines seem to indicate the desirability of programming both the item generation and test assembly procedures in a single integrated system for the purpose of extending the domains included in the test.

The new cloze item format has been applied to a variety of content domains relevant to the school setting to generate more than 13,000 test items. A sufficiently large number of items is now available to assemble domain-referenced tests of literal comprehension for a variety of evaluation purposes. The present chapter explored a number of applications of the testing materials in three widely-used forms of testing or evaluation.

---

[9] Actually, the domain-referenced model seems to imply little about meeting the conditions for construct validity. Messick (1975) indicates that this is no less a requirement for domain-referenced tests than for many other tests of ability and achievement used in education.

Presently the project has turned toward applying the cloze testing materials in a survey design in an urban district. The results of this application showed that, even at this stage where the testing materials are yet in a rudimentary, "paper-based" state, large numbers of test forms could be quickly assembled to mount a multi-matrix sampling design in a large population of readers with a broad range of reading levels. The data resulting from this test administration are currently under analysis both to permit refinement of the item format and to explore its validity. As chapters VIII and IX will show, the preliminary data seem to indicate that the item format does not yet require any major modification. In addition, there seems to be substantial indication that the overall test fits the theoretical and practical model constructed for it. Although these results can only be taken as preliminary or tentative, the promise offered by the testing model appears to have been justified to the extent that additional or expanded work on the testing materials is warranted.

CHAPTER VI

TEST VALIDATION AND REFINEMENT

Starting with an analysis of measurement needs in the area of reading, this report has proceeded by stages to describe the development, purpose, and characteristics of a new approach to assessing a basic level of reading comprehension--literal comprehension. The current state of this assessment system which, as noted, consists of some 1,374 cloze passages, associated items, and other testing materials, provides a broad foundation for the study of its validity. The approach planned for validating and refining the cloze format is a series of concurrent efforts designed both to study the meaning of the test and to bring the testing materials to a broadly usable state.

This chapter first provides an overview of the activities underway and planned for research and development on the cloze segment of the TDN. This plan includes the calibration of the test passages, studies of the validity of the cloze exercises as a test of achievement and ability, and research designed to make the test broadly usable in practice. The focus of the discussion from there is on the latter two topics, with most attention given to the topic of test validity. The problem of calibrating the test passages is treated at length in the next chapter.

Research and Development Overview

The overview of the research and development plan for the cloze exercises is shown in Figure 6.1. Examination of this figure shows that

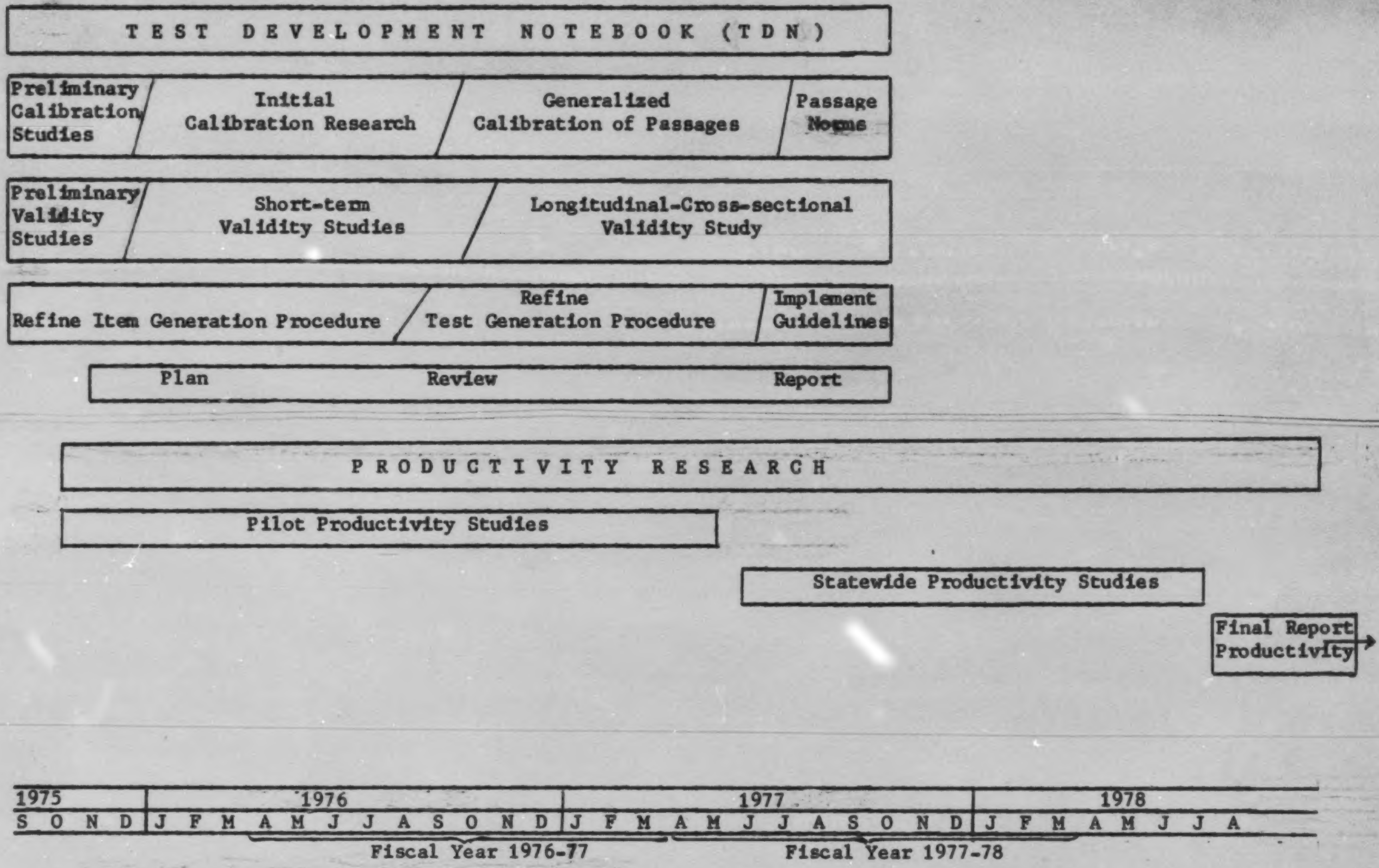


Figure 6.1. Research and development activities.

this research is embedded in a broader issue referred to as productivity research. The cloze and other segments of the TDN are being developed in part because of the need for improved measures of school output in studies of productivity in reading instruction. However, since the concern of this document is with the testing materials, the forthcoming discussion deals only with the lines of research activity projected for the TDN in Figure 6.1.

The first line of this research on the cloze exercises refers to the problem of calibrating the test passages on a single, equal-interval scale. As shown in Figure 6.1, preliminary calibration studies of the test passages have been underway for some time. Using the May-June test administration as a data source, a new computer program (Wright and Mead, 1975) was applied to determine the applicability of the Rasch model. The Rasch model appeared to accurately define the trait underlying the cloze exercises, and as a result, the additional stages of the calibration research shown in Figure 6.1 were justified. As described in Chapter VII, these additional stages project a further period of experimentation with the model, followed by a general application to the total pool of cloze passages.

The second line of research activity for the TDN, which is concerned with different types of test validity, was begun as a series of preliminary validity studies, based also on the May-June test administration. This initial effort, which continues to date, is expected to provide a basis for planning a second stage of test validation, identified in Figure 6.1 as "Short-Term Validity Studies." In these studies, a number of critical issues relating to alternate interpretations of cloze test scores, not all of which are necessarily identified to date, will be investigated with small student samples. Subsequently, the testing materials are again to be adjusted or the underlying concept realigned before mounting a large-scale study of the test using virtually the total population of an urban school district of 13,000 students. This larger study will combine the features of short-term, longitudinal, and cross-sectional studies in an intensive effort to further clarify the psycholinguistic meaning of cloze test scores across the grade 1-12 student population.

Concurrent with both the calibration and validation research, an effort will be initiated to program the current item and test generation procedures into a generally exportable routine that will be usable in various settings, such as state education departments, city districts, and regional institutions that coordinate complex technical educational services. In addition, workable models and guidelines will be drawn for using the test assembly procedure for supporting a variety of evaluation purposes, ranging from complex evaluation studies of reading programs to diagnosing the level of reading materials that an individual can comprehend literally.

Framework for Studies of Test Validity

As noted previously, work on test validity has been organized into the three broad lines of activity identified in Figure 6.1. All of these activities have to do with either establishing a basis for test validity or investigating the validity of the cloze test format. Before describing the specific research activities planned for examining the validity of the cloze exercises, the framework used for determining the kinds of validity studies deemed necessary is made apparent. Table 6.1 summarizes this framework.

The first type of validity referred to in Table 6.1 is content validity. Content validity focuses generally on demonstrating how well the test samples the classes of situations to which a test score is to generalize. Detailed

Table 6.1

Summary of Relevant Types of Validation[1]

Type of Validity	Sample Questions of Interest	Sample Types of Data/Analyses Needed	Judgements/Decisions Made
Content Validity	How well does the test sample the universe of responses and situations about which conclusions are to be drawn?	(1) Retrace the sampling of content. (2) Determine reproducibility of item generation. (3) Determine agreement on content categories sampled. (4) Determine reproducibility of test generation.	(1) Judge degree of bias in content representation. (2) Identify subjectivity in procedures used to generate items or tests. (3) Adjust content representation in the test.
Educational Importance	Does the test measure an important educational outcome?	(1) Demonstrate relationship of test to educational objectives. (2) Demonstrate that the characteristic measured by the test is valued and of practical importance in a variety of situations.	(1) Test is considered a relevant measure of outcomes. (2) Potential social and educational utility is determined.
Construct Validity	Does the test measure what it purports to measure? Is the characteristic measured by the test one that is influenced by the educational process? What is the meaning and interpretation of a test score? What are the educational and social consequences of using the test?	(1) Studies of convergent and discriminant validity. (2) Relationship of test score to amount of schooling and short-term interventions. (3) Developmental studies of consistency across populations and domains. (4) Studies of response processes by age. (5) Studies of dimensionality.	(1) Determine possible meanings and uses of test scores in practical situations. (2) Modify the test to improve consistency with construct and interpretations. (3) Modify and extend the meanings surrounding the test.
Placement	Is performance improved when students are assigned to instructional materials or conditions on the basis of test scores?	(1) Assign students to instructional levels and ranges of materials using test scores and compare to unassigned students.	(1) Determine ability to generalize from a test score to instructional and other reading contexts.

---

[1] Adapted from Cronbach (1971, p. 446).

evidence of content validity is of particular relevance for the cloze testing materials since it is assumed that a score on any test assembled from the testing system can be interpreted directly in terms of a person's ability to read in a specific universe of written discourse. In establishing content validity of a domain-referenced test, the investigator must demonstrate that the test accurately samples the domains to which the test is intended to generalize. Also of concern here is the adequacy of the universe definition and the objectivity or reproducibility of item construction.

Content validity is established largely by empirical means, for example, by referring the content sampling plan to test users. The importance of what is measured by the proposed test is established largely in the theoretical statement which defines the construct underlying the test. That is, the definition of what the test purports to measure not only establishes the theoretical importance of the test, but also explores the social and practical implications of test use (Messick, 1975). Evidence of the importance of a test is thus initially a problem of logic, coherence, and the adequacy of the construct definition, but it is also ultimately determined by empirical results which reflect negatively or positively on the network of concepts defining the test and its uses. The importance of the cloze segment of the TDN seems to be adequately established in the construct definition, but still to be shown is evidence that the test accounts for something educationally and psychologically meaningful.

The third type of validity relevant to planning the course of research on the cloze format is construct validity. Because the cloze exercises claim to measure a particular type of comprehension and because this claim has a number of very important implications from theoretical, decision-making, social, and policy points of view, it is critical that the psycholinguistic and practical implications of scores from the test be explored and established (Cronbach, 1971; Messick, 1975). In educational measurement, there has been a tendency to regard an achievement test as terminally valid if its importance (measures recognized objectives) and content validity are established. Or, a new test of achievement may be validated against several established tests, whose validity ultimately also rests on older claims of importance and content relevance. However, Messick (1975, p. 956) has pointed out that:

> . . . even for purposes of applied decision making, reliance upon criterion validity or content coverage is not enough. . . . the meaning of the measure must also be pondered in order to evaluate responsibly the possible consequences of the proposed use.

For the cloze segment, this view requires a series of interrelated studies, some involving investigations of convergent and discriminant validity, some relating to the consistency of the test across populations and situations, and still others involving examination of the process of responding to the test or the effects of instructional interventions--to name a few. Examining the construct validity of the cloze test, it will be seen, consumes the larger part of the present validation effort.

The last type of validity identified in Table 6.1 refers to the utility and accuracy of the test in instructional decision making. The cloze testing materials are intended to provide a basis for a variety of educational placement-type decisions, such as determining the degree of fit of texts in a given content area with the comprehension abilities of groups of students or assigning texts in a particular range of readability to an individual student.

Studies of Test Validity

Continuing or planned studies of the validity of the cloze testing materials are reported below in the organization presented previously in Figure 6.1. The types of validity relevant to establishing the cloze format defined in Table 6.1 are reflected throughout this report, with the exception of the factor of importance, which, it is felt, was substantially established in the theoretical discussion of literal comprehension in chapters II, III, and IV.

Preliminary Validity Studies

Two avenues of investigation of test validity were initiated in the preliminary phases of research on the cloze test. The major part of the preliminary effort is based on the May-June test administration and has largely to do with refining the test and tentatively exploring its construct validity. The second avenue of the research constitutes the beginnings of a series of content validity studies. Each of these investigations is discussed in turn.

Construct validity. The organization of the data collection for this component of the preliminary phase of the validation effort was presented in the latter part of Chapter V. To recapitulate briefly, this study consisted largely of gathering data from the administration of three types of tests in a grade 1-9 population of 5,000 students: (a) a multiple-choice cloze test; (b) a multiple-choice comprehension test composed of modified wh-items; and (c) a standardized achievement test given annually by the study district. The cloze and wh-tests were initially conceived as two different measures of the same construct of literal comprehension, with the latter test having been constructed because an adequate, alternate measure of the construct was not available. The standardized test used in the district, which was the California Achievement Test or CAT (Form A), contained measures which potentially converged or diverged with the concept of literal comprehension measured by the cloze test. A list of the variables defined by these tests is given in Table 6.2.

Table 6.2

List of Variables Measured by Tests Included in the Preliminary Validity Study

Cloze Test	Wh-Test	California Achievement Test[a]
Total Cloze Score	Total Wh-score	Vocabulary
Cloze Paragraph Scores	Passage Independence Score	Letter recognition
Noun Score	How score	Word forms
Verb Score	What score (noun, pronoun)	Word recognition
Adverb Score	What score (verb)	Picture-word association
Adjective Score	When score	Words in context
	Where score	General Comprehension
	Which score	Locate facts
	Who score	Interpretation
	Why score	Relationships
		Generalizations
		Draw inferences
		Comprehension/Social Studies
		Comprehension/Science
		Comprehension/Mathematics
		Language Skills
		Sentence structure
		Transformations
		Mechanics
		Usage
		Verbal IQ
		Non-verbal IQ

[a] Not all subscores listed are available for each test population.

This initial test administration had several purposes. The first objective was to combine logical analysis of the consistency of application of the multiple-choice cloze item form with conventional and Rasch item analysis data with the intent of conducting a first refinement of the total item pool. As expected, this activity led to a number of changes in the rule system for selecting and processing a passage in the multiple-choice cloze format. This activity further resulted in a major revision of the item pool which ultimately affected an estimated 85% of the 1,374 passages already on paper-punch tape.

A second major purpose involved use of the Rasch item statistics in determining the adequacy of a unidimensional model in accounting for the hypothetical underlying trait of literal comprehension across so many different cloze test forms (N = 36) and populations. In this activity, the distributions of item difficulty (more appropriately called item easiness in the Rasch model output) and ability were also examined in detail to determine the extent to which the testing system was consistent with the domain-referenced model.

A third major purpose involved examining the internal consistency of the cloze test and the wh-item format through conventional item analysis techniques. In this effort, the various part scores of each test type were intercorrelated and the Kuder-Richardson Formula-20 reliability coefficient was calculated on the total test score on each test form (N = 72 test forms). In addition, for the cloze exercises, the correlations of noun, verb, adjective, and adverb subscores with total test scores were calculated, corrected for the correlation of each part score with itself in the total test scores (N = 36 test forms). Together, these analyses reflected the consistency and uniqueness of the four types of deletions made in the cloze item form.

Finally, in an attempt to examine convergent and discriminant validity, the various subscores of all three types of test were intercorrelated for each CAT test level population (Level I, II, III, and IV). These analyses were designed to yield a set of validity coefficients which could be examined for consistency with theory. This analysis remains very tentative at this point due to the difficulties involved in accurately expressing the psycholinguistic meaning of what is measured by the various items included in the CAT. Because of arguments raised in Chapter II, strong confidence cannot be placed in the tests based on the wh-item format either, particularly as used here, where the test was given across such a wide age range.

Currently, the intercorrelations of the subscores in the three types of tests available for this phase of the investigation are being subjected to a principal components analysis and verimax rotation for each level of the sample. The results of this effort will be made available at a later point in time.

Content validation. Work in this area was begun in the fall of 1975 with the selection of a regional, representative sample of 192 school districts. The cooperation of the individual school districts in this sample has now been secured, and in early 1976, each district will receive first a lengthy questionnaire and instructions. This questionnaire will include a set of labels defining major clusters of the objectives of reading instruction. In addition, the respondent will receive a set of scaled passages chosen from the wh-question pool to represent readability levels 1-20. The respondent is to indicate the clusters of objectives taught in reading instruction in grades 1-6, their levels of emphasis in instruction, and the associated ranges of passage difficulty for the reading materials used in instruction with each grade level population.

In part, the study is intended to define the levels of passage difficulty that students in various grade levels are routinely expected or believed to master in school, i.e., are presumed able to comprehend at the literal and higher levels of reading comprehension. The school sample is well defined, thus enabling the partitioning of the findings by region; by urban, suburban, and rural districts; and by characteristics of the student population.[1]

This initial effort at content validation is expected to provide a basis for defining the ranges of passage difficulty reflected in the instructional experiences of part of the student population for which the cloze testing materials are intended. These data will become part of a more extensive effort to define the content domains of reading instruction from different perspectives.

During the spring of 1976, this same study sample is expected to respond to a second questionnaire which will attempt to obtain a more representative definition of the content categories that express the most frequent areas of written discourse experienced by grade 1-12 students in the school and extra-school environments. Based on the results of this study, the content of the cloze component of the TDN will be further adjusted and a more definitive effort will be made to describe the readability characteristics of types of written discourse by grade level or in relation to other relevant situations. The size of each major universe of written discourse identified as important to a level of the 1-12 student population will be estimated. For each such stratum, and using a 98% confidence level, readability samples of 100 word passages will be drawn. Subsequently, the readability levels of each sample will be computer analyzed using the Dale-Chall index, the resultant distributions of readability will be arrayed, and means and standard deviations calculated.

---

[1] The study will also contrast the readability ranges reported by teachers with the readability ranges of the passages included on five major standardized tests of reading comprehension.

The results of the foregoing analyses are expected to provide an adequate basis for adjusting the content of the cloze component of the TDN prior to any serious effort at calibration or more extensive validation. In addition, the information on expected and actual readability levels of relevant areas of written discourse will provide a basis for defining a normative context for the interpretation of test scores resulting from the administration of the cloze format. The sample distributions of readability levels of written discourse in various categories are required to provide a basis for generalizing from a test score to a domain of written discourse.

Short-Term Validity Studies

Short-term validity studies, referred to previously in Figure 6.1, constitute a group of related efforts designed to address some crucial issues of test score interpretation. The results of these studies will guide the revision of the cloze exercises in preparation for further calibration studies and longer-term, more expensive efforts at test validation. Table 6.3 offers a framework for organizing these studies in terms of four groups of variables in the test situation that may affect the interpretation of a test score based on the existing cloze format. The first group of variables refers to important global characteristics of the types of written discourse now included in the test, variations of which may cause the test to measure other than the hypothesized literal comprehension factor. The test format is presumed to be an effective mode for measuring literal comprehension that is essentially invariant across written material that differs in terms of specific content area, semantic complexity, or syntactic complexity. Contextual constraint refers to a specific problem recognized in processing passages into the multiple-choice cloze format. This is the issue of whether the meaning of the test score varies as a

Table 6.3

Framework for Organizing Short-Term Validity Studies

Relevant Sources of Variation in the Test Situation			
Passage[a]	Response	Context	Person
Contextual constraint	Semantic competition	Training	Personality
Syntactic complexity	Syntactic competition	Instructions	Oculomotor skills
Semantic complexity	Content words	Time	Orthographic competence
Content area	Idioms, metaphors	Examiner	Phonological competence
			Semantic competence
			Syntactic competence
			Specific knowledge
			Speed
			Memory
			Verbal reasoning
			Non-verbal reasoning
			Comprehension
			Literal
			Non-Literal
			Test-wiseness

[a] Passage factors are represented by a multitude of variables. For example, the semantic component includes vocabulary load, metaphorical usage, level of abstraction, etc.

function of the relative immediacy of the passage context that in turn determines the response to a given deletion in a passage. This context may be intra-sentential, intersentential, or extra-sentential, depending on how the rules for deletion are applied.

The preliminary validity studies have to date provided a holistic evaluation of the combined effects of content area, syntax, and semantic factors on the interpretation of cloze test scores, since all of these factors are indirectly if somewhat inefficiently measured by application of the Dale-Chall index to the test passages. Additional studies where these factors are examined in isolation in terms of their effects on test score interpretation are planned for the spring and summer of 1976. It is intended that the issue of the effects of contextual constraint on cloze test scores will also be studied at that time.

The general approach to the studies of passage effects will be to vary passages on the particular dimension of interest, such as syntactic complexity, while holding other dimensions constant. The contextual conditions within passages that appear to primarily determine the response process will also be categorized separately. Correlational analysis may then be used to analyze the contributions of selected person factors (column 4 in Table 6.3) to item and passage responses as a basis for inferring potential changes in the meaning of a cloze test score produced by one or more passage variables. This approach to analysis should provide important leads concerning the passage variations that interact with cloze deletion rules and that may cause the test to unduly emphasize general reasoning or other factors as opposed to those factors which are presumed to contribute to a literal understanding of the meaning of a given sample of written discourse in a level of the student population.[2]

A second group of studies is planned for spring and summer, 1976, to address unresolved problems surrounding the preparation of distractors for a cloze passage (Table 6.3, column 2). According to the theory surrounding the test, semantic competition among distractors will unduly emphasize knowledge and reasoning skills which are essentially extraneous to the literal comprehension of the passage and which may be synonymous with other more complex measures of comprehension or intelligence. This interpretation needs to be evaluated in different strata of the population with both semantically interfering and non-interfering distractors applied to the same passages. Similarly, the adequacy of using grammatical class as a basis for selecting distractors, and thus presumably controlling for syntactical competence alone being the basis for selecting the correct answer on the cloze test, needs to be examined before the computerized distractor generator can be reasonably finalized.

---

[2] What might be expected is that syntactic or semantic competence are emphasized at varying levels of complexity by a passage or particular deletion depending on its organization, content, and structural complexity.

A final issue of this type, identified last in column 2 of Table 6.3, will be examined in a brief study designed to determine the possible need for modifying the distractor generation process when the word deleted from a passage is part of an idiom, or a metaphor, or is a specific subject matter word. Presently, the rules governing the generation of distractors for these types of deletions are simple and straightforward. The American Heritage Word Frequency Book (Carroll et al., 1971) is used to identify specific content words in a passage and metaphorical and idiomatic language are largely ignored as a basis for modifying the rules for distractor generation.[3]

The third group of short-term studies that is required before the cloze format can be effectively modified for large-scale validation and calibration is identified in the third column of Table 6.3 as the testing context. The amount of training needed to standardize the test across a widely ranging student population is an important research issue because the test uses a new, unfamiliar format. The related matters of test instructions and examiner behavior will also be examined in applied research which will attempt to determine the conditions under which test-taking motivation may be maximized while minimizing the effects of guessing on the test score. Finally, this element of the research effort will attempt to define optimal amounts of time for test-taking. The intent of this area of the research will be to produce an examiner and examinee training package that will orient the test-taking situation to one that is more psychologically attuned to the taking of a domain-referenced test.

---

[3] Under some conditions, a word that forms part of an idiom may not be a candidate for deletion (See the rules for generation of distractors in Appendix A).

In concluding this brief presentation of the short-term phase of test validation, it seems necessary to point out that the foregoing set of planned studies represents only some of the larger issues of validity recognized to date in work on the cloze testing materials. In practice, the present approach to investigating the effects of a particular variable on what the cloze format measures is likely to be relatively holistic. For example, in actually measuring semantic complexity, it may be necessary to focus on traditional approaches based largely on vocabulary load. Semantic complexity has a potentially large number of referents and dimensions and, as an area of investigation relating to the cloze process, could easily consume the whole of the resources devoted to this research. Wherever possible, the tendency will be to use an existing measure. The exceptions to this approach will be in the work described in the next section which will attempt to create an alternate and less ambiguous criterion for measuring literal comprehension as well as mount a parallel effort to create measures of syntactic competence and syntactic complexity.

A Longitudinal, Cross-Sectional Validation

A one-year study combining the elements of short-term, longitudinal, and cross-sectional designs will constitute the basis for a broader examination of the construct validity of the cloze testing materials. This phase of the research will attempt to address some larger issues concerning the validity of the multiple-choice cloze format, while continuing to provide a context for examining research issues previously identified. That is, detailed analysis of items, passages, and item format variations will continue here and even be extended to additional content areas. However, the focus of this phase of the research will be on the developmental and instructional implications of the construct of literal comprehension. Sample questions of interest will be: To what extent do literal comprehension scores change in relation to the passage of instructional time over 12 grades? Is the development of literal comprehension, as measured by the test, continuous over the school years? Or, does more development occur in some years than in others? Which students develop the skill most rapidly? To what extent is the development of literal comprehension influenced by manipulable home and school factors? By non-manipulable home and school factors? And finally, to what extent does the development of literal comprehension affect other school learning tasks? Does literal comprehension contribute to academic and personality development in the school?

Study sample. The proposed study is to take place in the same urban district that provided the data for the preliminary phase of the research, except that it is expected that virtually the total student body will participate in the longitudinal study. This will provide a heterogeneous sample with a size of more than 1,000 students per grade level at each of grade levels 1-12. The idea of conducting the study in this single district contributes substantially to the economy and feasibility of the research, while, it is felt, not greatly affecting the ability to generalize results. The composition of the student body is fairly representative of the major cultural and economic strata of the New York State population, except that minority elements tend to be somewhat overrepresented. The district has an excellent standardized testing program, keeps accurate and complete records on the student population, and is very supportive of research. There is considerable positive communication and interaction between the district and the larger community, thus creating the conditions that will be needed to obtain the parent interview and questionnaire data to be collected in the study. Finally, because of substantial special Federal and State monies annually infused into the district, there is extensive variation both between and within schools in the amount of time and resources devoted to reading instruction. This variation will provide part of the background for assessing the relative effects of school factors on cloze test scores as contrasted with their effects on other measures of reading.

General study design. The categories of measures to be collected in the study are listed in Table 6.4. This is essentially an expansion of Table 6.3 to include measures of school and home factors. Selected personality and achievement test factors have also been added to the "person" column. In large part, adequate measures of these factors are either available from the study district's testing program or from previous research of this type (cf. Kidder et al., 1975). The new instrumentation that will be constructed for the study includes a measure of syntactic complexity applicable to reading materials, a measure of syntactic competence, applicable to the student population, alternative measures of literal comprehension, and questionnaire and interview schedules that will be used to determine the breadth and complexity of the reading experiences of the student population after the manner described by Chomsky (1972). The selection and organization of variables and measures, as shown in Table 6.4, provides a basis for defining the syntactic and semantic components and

Table 6.4

Framework for Organizing Variable Measures in the One-Year Validation Study

Passage	Response	Person	School	Home
Syntactic complexity[b]	Content words	Personality	Complexity reading materials	Socioeconomic status
Semantic complexity	Idioms, metaphors	Anxiety	Instructional time	Cultural background
Content area		School satisfaction	Instructional mode	Parent-child reading experiences[b]
		Self esteem	Learning environment	Child's reading experiences[b]
		Orthographic competence[a]	Teacher age	
		Phonological competence[a]	Teacher experience	
		Syntactic competence[b]		
		Semantic competence[a]		
		Word knowledge		
		Specific knowledge[a]		
		Mathematics achievement		
		Science achievement		
		Social Studies achievement		
		Language Arts achievement		
		Verbal reasoning[a]		
		Non-verbal reasoning[a]		
		Comprehension		
		Literal[b]		
		Non-literal factors[a]		
		Sex		
		Age (mos.)		
		Grade level		

---

[a] Available from the district standardized achievement testing programs.

[b] To be constructed for the study.

levels of the major dependent variable of interest (literal comprehension) while attempting to trace the contributions of immediate and contemporary antecedents of reading performance in the home and school.

The procedures for assembling the primary measures of literal comprehension to be used in the study will initially involve the selection of some 600 cloze passages from the TDN distributed equally across the total range of readability levels and in the proportions of 2/6 for the reading/literature category and 1/6 for each of the additional textual categories.[4] These passages will provide the raw material for the calibration pilot described in Chapter VII, with the result that this pool of passages will be calibrated on a common, equal-interval scale.

Subsequently, the scale for the reading/literature stratum of 200 passages will be divided into 30 equal intervals. Each successive set of 6 such intervals will constitute a test level, with no overlap between levels. Stratified random sampling will then be applied to the passages in the intervals in each test level to obtain 6 parallel test forms of 6 passages each per test level and a total of 30 forms across test levels. The number of passages required for this design is 180 (36 passages x 5 test levels).

The foregoing design will be repeated for each additional textual area, except that the design parameters are 100 passages from which 3 passages will be sampled from each interval in a test level. This design will require 90 passages and will result in 3 parallel test forms of 6 passages per test level.

A special placement test will then be constructed from the reading/literature passages for the purpose of estimating the test levels of individual students. Thereafter, the student population will be assigned to test forms by test levels within grade levels using random stratified sampling to insure equal numbers of students across test forms. This process will be repeated for the test forms in each additional textual area, using each time a different randomly constituted one-fourth of the student population at each grade. Students in each grade level population will thus receive two parallel cloze test forms on each of three test occasions during the school year (one in reading/literature, one in a content area), with a different set of forms used on each test occasion. The outlines of this design are shown in Figure 6.3 by identifying each test occasion with a C.

---

[4] It will be necessary to add some new passages at the upper ends of the scale for the 11th and 12th grades.

At least two additional testing sessions will be required of each student in the study sample. At the pretest and posttest occasions identified by W's in Figure 6.3, each student will receive on each occasion a different test form that constitutes an alternate measure of literal comprehension. These test forms will be assembled from the wh-item pool using a design similar to that described in Chapter V, except that the passage independent section of the test will be replaced by main idea and title items. (For each of the 300 wh-item passages, there are available up to 4 verbatim or derived main idea and title items). Additional alternative measures of literal comprehension based on the paraphrase transformation and interviews will also be administered to small subsamples of the study population.

The foregoing design will require approximately 120 minutes of testing time per student on the pretest and post test occasions and 80 minutes at the interim data point. This element of the study design will provide a basis for estimating the mean and variance of literal comprehension change

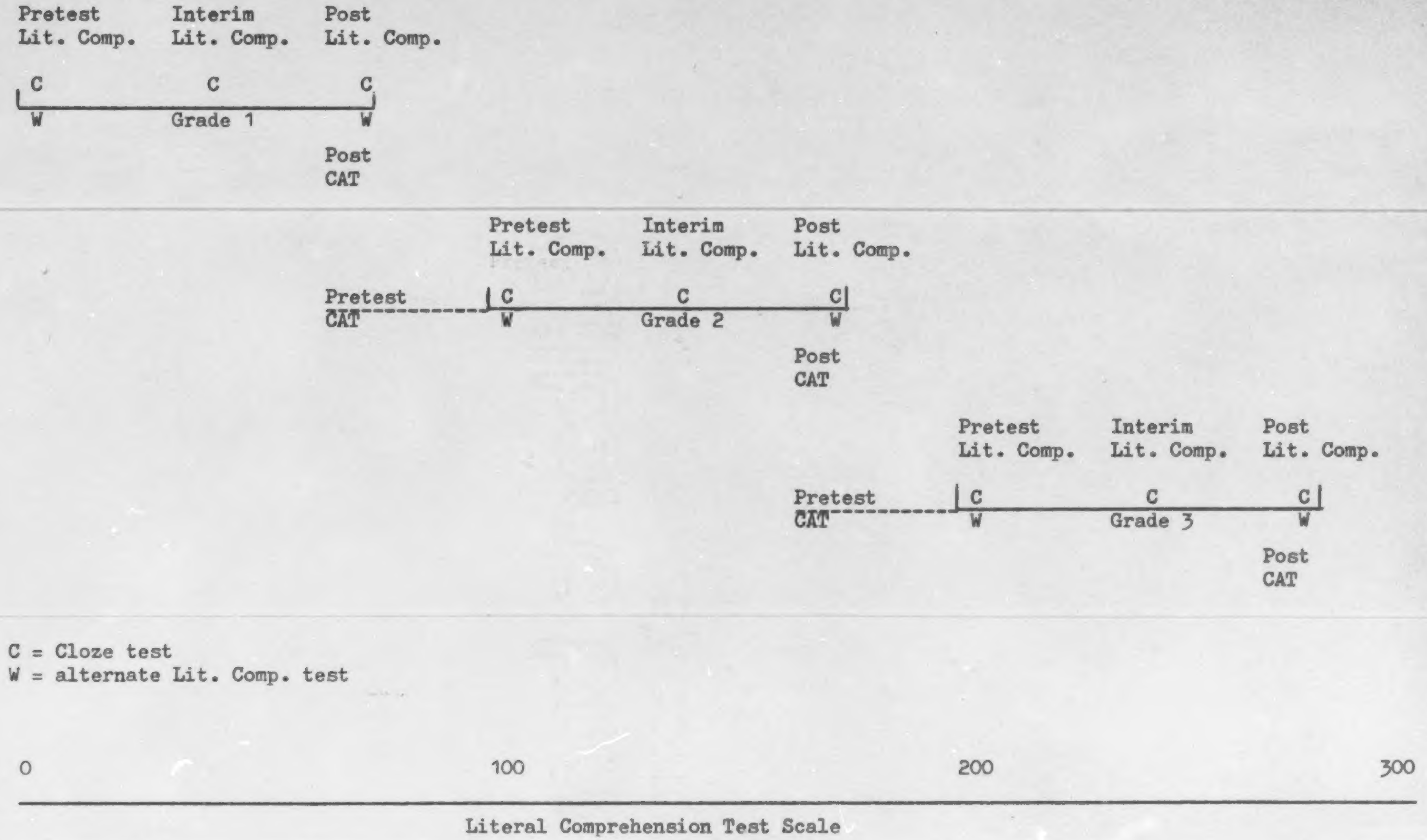

Pretest
Lit. Comp.
Interim
Lit. Comp.
Post
Lit. Comp.
C
C
C
W
Grade 1
W
Post
CAT
Pretest
Lit. Comp.
Interim
Lit. Comp.
Post
Lit. Comp.
Pretest
CAT
C
C
C
W
Grade 2
W
Post
CAT
Pretest
Lit. Comp.
Interim
Lit. Comp.
Post
Lit. Comp.
Pretest
CAT
C
C
C
W
Grade 3
W
Post
CAT
C = Cloze test
W = alternate Lit. Comp. test
0
100
200
300
Literal Comprehension Test Scale

Figure 6.3. Longitudinal and cross-sectional components of the test administration for major dependent variables in the one-year validation study.

scores across the school year by grade levels. Because all such estimates are referenced to the same scale, it should prove possible to estimate the total possible change across the scale by suitably combining cross-sections or grade levels. The variation in change scores within grade levels may also be partitioned to determine the contributions of school and family factors to the development of literal comprehension. This same analytical process can also be followed for the standardized reading test scores available on the grade 2-9 student population where pretest and posttest scores will be available for each of the populations in these grade levels.[5]

Many of the details of the foregoing design yet need to be worked out. Some elements of the design must be restricted to small cross-sections of the student population due to the costs of developing adequate instrumentation. For example, the measures of syntactic complexity and competence will very likely be restricted to the elementary grades because of the necessity of validating the measures to be used (Athey, 1975; Finn, 1975).[6] However, with these exceptions, it is expected that the basic outlines of this design can be installed with the result that data will be available on many interesting questions about the psychological meaning of the construct, literal comprehension, and of some of its operationalizations, including the cloze format. Because of some of the unique features of this design, e.g., common item formats administered at each level of the population and a single underlying scale to which test scores can be referenced, the results of this study should be of broad interest to the professional reading community.

---

[5] The CAT offers an ADSS scale score which allows the combination of test forms and levels on one scale, thus making the analysis possible from a psychometric point of view. However, unlike the cloze test, the content and format of the CAT changes radically from level to level. It should prove interesting, however, to determine how much the student population changes on this scale and how much of this change is associated with school influences.

[6] Personal communications on this issue indicate extensive costs for developing instrumentation across grades 1-12. However, some models exist for use in grades 1-6.

Generation and Use of Items and Tests

Concurrent with the course of test validation, identified in Figure 6.1 at the outset of this chapter, efforts will continue to improve the reproducibility and exportability of the item and test generation procedures. Also, this section of the research program will produce material that will help potential users apply the cloze testing materials to broad evaluation problems, ranging from individual measurement problems to large-scale evaluations of reading programs within and between school districts. This final section of the present chapter briefly discusses the overall content of each of these efforts.

Test Item Generation

The test item generation process is now largely a continuous, human operation that begins with the sampling of passages from sources and concludes with a finished passage in the multiple-choice format. Aside from the sampling of original passages, it appears that this entire routine can be computer programmed in the interactive mode, with the laborious and repetitive components of item generation handled entirely by the computer. For example, following the sampling of a passage, the computer would analyze passage readability and identify various potential deletion patterns while also identifying the characteristics of each deletion pattern (percent of text deleted, percent of nouns deleted, etc.). The person interacting with the computer would then indicate a particular pattern of deletion, would further indicate the word lists to be accessed for each deletion, and would finally be presented with a tentative clozed passage. The resultant cloze passage would be inspected for any departures from the deletion rules, word lists would be further accessed as needed, and, finally, the completed passage would be programmed for type style, letter size, layout, etc.

This projected automation of the item generation routine will make the process of test item development generally exportable. This activity must be carefully integrated with the research planned on the validity of the item format.

Test Assembly

The test assembly process is yet relatively crude, consisting largely of assembling the passages relevant to a given evaluation design into tests and delivering the associated paper-punched tapes to the printer. Over the course of the research, it is expected that the test assembly process will also be programmed in the interactive mode to operate on the passage pool and enable an evaluator or researcher to assemble a test or tests for a particular purpose.

This process first requires that the pool of passages exist in a computer file, along with the data that become the selection criteria of the user. Relevant data for passages include Rasch calibrations, readability characteristics, content area, psychometric characteristics, and so on. The process of passage selection must allow the user to apply several selection criteria simultaneously, while also providing for various sampling strategies. The programming should be sufficiently sophisticated so that a test or tests can be generated with predictable evaluative and psychometric characteristics. For example, the program should be able to deliver $\underline{n}$ parallel tests, with specified content, of fixed length; and with a projected mean, standard deviation, and reliability in the population.

This section of the research program will also provide a set of practical models or guidelines for using the cloze component of the TDN in evaluation and research. Of particular interest is the derivation and effective presentation of workable applications of matrix sampling, using the cloze passages as a resource. Sirotnick (1974) attempted to provide such models in a general presentation, but here it would be expected that a number of detailed models could be derived and applied in simulated use of the cloze passage pool. Sirotnick's presentation showed how a school district could save considerable amounts of testing time and money by applying matrix sampling to a large number of items and domains of content. The flexibility of the cloze component was planned largely so that the evaluator could begin to take advantage of the economy and efficiency of the matrix sampling model in evaluation in reading.

Presumably, in the finished product from this phase of the research, the user would first explore the simulated evaluation models projected here. He would then specify the parameters of the evaluation design that fitted his situation (e.g., number of test groups, numbers of tests, sampling plan for each test, confidence levels, etc.). Armed with these parameters, he could then use the test assembly program to generate the required tests in paper-punched tape form.

Decision-Making Utility of the Test

This component of the practical side of the research will go beyond the processes of assembling items and tests to show the user how the test data may be used in certain types of practical decision making. The decisions to be addressed are largely of the placement type. The focus is on assigning an individual or group to reading materials or to levels of the curriculum. The basic problem to be addressed involves creating the technical guidelines and techniques that will allow a user to generalize from a cloze test score (in Rasch calibrations) to a segment of one or more domain distributions of readability. The problem is illustrated in part in Figure 6.4. For each grade level, there is a distribution of readability of the written material in each content domain and subcategory. Generally, the user wants to know how to assign a student to sections of a set of such distributions so that the student can be given material he can actually read.

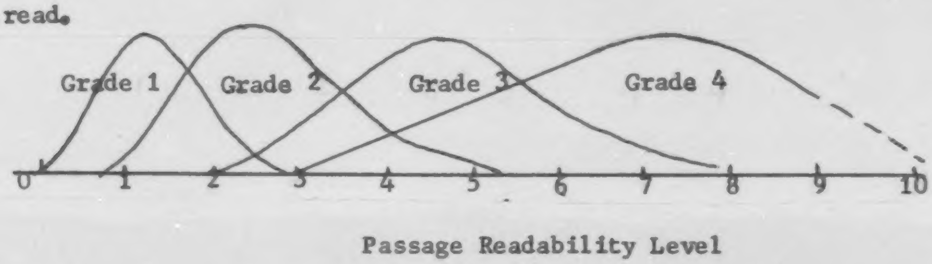


Figure 6.4 Hypothetical distribution of passage readability for one content area by grade level

As part of the research, the sampling distributions of readability scores of the domains and subcategories of content included in the cloze test will be known, as will the distributions of readability scores for passages with the same Rasch calibration. The problem is to work out the technical parameters of using both sets of data to predict the range of readability that is appropriate for an individual with a given score on the cloze test. By generalization, the resultant model should further show how well any given group "fits" the domains and categories of reading materials sampled by the test.[7]

In summary, this aspect of the applied component of the research will derive practical procedures for estimating the probable level of functioning (in terms of literal comprehension) of an individual or group in relation to a given domain or category of written discourse. It would also be of considerable value, if resources allowed, to extend the model to show a local user how he could take into account the readability characteristics of his own populations of students and reading materials to achieve a generally improved match between student abilities and instructional materials.

---

[7] Included here is the problem of defining statistically the cutoffs that indicate the point of comprehension/no comprehension in a student's test protocol. These cutoffs may vary by age or grade level.

Conclusion

The foregoing discussion is an expansive outline of some of the studies that might be conducted to examine the validity of the multiple-choice cloze format as a measure of literal comprehension. The projected studies are cast in the scientific framework of construct validity. That is, because of the implications of meaning and use attached to the cloze testing materials, it is deemed necessary to embed the test in a research program that will tend to clarify both the meaning of the concept underlying the test and the utility of the test in measuring the concept. Typically, content-oriented or achievement tests are not embedded in such a research program, but are considered valid on the grounds of content and convergent validity.

The validation program defined here projects the gathering of evidence of different types of validity on the cloze test. The content validation phase is a much more extensive effort than is usually mounted with tests of achievement. Here, the research program will attempt to characterize the content domains to which the test is referenced in considerable detail. This effort will ultimately provide an improved basis for using a reading test score, based on the multiple-choice cloze, in decisions that directly affect how individuals and groups are assigned to domains of reading instruction.

The construct validation phases of the proposed research are concerned with evaluating the main effects and interactions of organismic and situational variables on the development of literal comprehension, as measured by the cloze exercises. The quasi-longitudinal component of this research on test validity further provides a basis for examining the theoretical and practical importance of the construct of literal comprehension--and the cloze exercises as an effective measure of the construct--across the years of public schooling. The potential outcomes of this research would appear to provide the kind of information that is needed to determine the utility of the cloze exercises as an important measure of output in reading instruction.

Finally, concurrent with the validity studies planned for the cloze testing materials, a program will be conducted to gradually transform the test into a state of practical utility. This program will make extensive use of modern technological developments with the intent of improving the economy, usability, and applicability of the testing materials. These applications constitute a model for future tests, which will not be tests in the usual sense, but devices that can be tailored to measurement/evaluation situations as needed.

CHAPTER VII

CALIBRATION AND SCALING DESIGN

In addition to justifiable construct validity, the new multiple-choice cloze testing system will have a very useful scale for score interpretation. The new test scale will have distinct advantages over the scales of commercially available, standardized tests. The scale for the multiple-choice cloze (MCC) testing system will have equal intervals and a "low-difficulty point" near zero. These properties alone will support the legitimate assessment of literal comprehension over time while permitting the use of unique test forms at each point in time. The reading passages comprising this scale will be drawn from eleven content domains and cover difficulty levels from first grade to college. The approach to be used in scaling these reading passages will allow the estimation of reading ability in all 11 domains from one test in one domain. Thus, the most useful application of the scale may be the construction of tests that are tailored to individual students. A teacher, working with the Test Development Notebook, will be able to select passages that will be targeted around an individual's true ability. A test so designed will provide a precise assessment of a particular student's ability in literal comprehension.

Actually, a teacher using the Test Development Notebook could construct unique tests for each student in a classroom, several times throughout a course of instruction or an academic year. This would provide a design for achievement monitoring that is seldom used in schools today.

Trait Definition

The first step in the development and calibration of a test scale is the specification of the trait under investigation. That is, what student trait is actually measured by the MCC test?

In the early phases of trait definition, singular operational definitions are counter-productive, "for the closure that strict definition consists in is not a precondition of scientific inquiry but its culmination" (Kaplan, 1964, p. 77). The veracity of the trait-definition being investigated, namely, literal comprehension, is open to strong criticism, especially when conceived of operationally (see Chapter II in this proposal).

Operationally, we might say that literal comprehension is exactly what the MCC test measures. Unfortunately, one inspection of a multiple-choice cloze test could result in several different interpretations of what the test might measure. In addition, if only an operational definition is provided, then the burden of construct validation is thrust upon the consumer, "who will inevitably make inferences beyond the universe of situations representatively sampled by the test" (Cronbach, 1971, p. 483). Thus, trait definition is tied directly to construct validation because users will demand, and legitimately so, that the test, if properly used, measure what the developers say it measures.

In order to maintain the desired interpretations of the MCC test, construct validation (see Chapter VI) must be designed carefully in order to refute any substantive counter-interpretations of the test that might arise from its use. This construct validation is actually a clarification and justification of the operationalization chosen for measuring literal comprehension. In order to be received properly, however, a particular operationalization should have a firm conceptual basis, especially in the behavioral sciences. The conceptual basis for the trait measured by the multiple-choice cloze test is explained in detail in Chapters II through IV of the present proposal.

Measurement Issues and Model

There are several measurement issues involved in the calibration of the MCC test. These measurement issues arise from two sources: (1) the MCC test format and (2) the requirements of the measurement model used to calibrate the test. Generally speaking, calibration means estimating item difficulties so that items can be scaled from the least difficult to the most difficult. However, in the present context, the emphasis must shift to passage calibration.

The format of the MCC test is radically different from that of conventional tests of literal comprehension. As described earlier in this proposal, no formal questions are asked in the MCC test. The student is simply required to choose from three, four, or five alternatives that word that has been deleted from the paragraph in question. The student's ability to reconstruct the original paragraph reflects apprehension of the literal meaning of the paragraph. The manifestation of this trait is considered to be an all-or-none phenomenon, that is, apprehension occurs or does not on a specific passage. Thus, the test format remains essentially the same from grade 1 through college.

The multiple-choice cloze format with no formal questions will reduce the importance of general intellectual skills in the student's response. The format is designed to measure literal comprehension of a passage, not the student's ability to comprehend and answer questions following the passage. In the latter instance, skills beyond literal comprehension are called into play. Thus, a major research question that must be answered is whether or not the MCC test is unidimensional and thus measures a stable trait across time.

A second complicating factor in the MCC test was the choice to delete only nouns, verbs, adjectives, and adverbs. The effect of this decision on the performance characteristics of the MCC test must also be investigated. This choice may complicate the attempt to maintain a unidimensional measure of literal comprehension due to a lack of systematic variation among the deleted words.

The Rasch measurement model will be used to analyze and calibrate the MCC test. This model has been chosen for two major reasons. When tests have been constructed so as to meet certain specifications, "application of the Rasch model gives person-free item calibrations and item-free person measurements" (Wright and Mead, 1975, p. 2). Such objectivity in measurement is seldom attained in the behavioral sciences. For example, if you want to know a person's height, you measure him with a yardstick or another device. Within reason, two different yardsticks will provide the same estimated height. What happens when students are given two reading tests designed by separate companies? Does one consistently get the same estimate of a student's reading ability with the separate tests?

> The Rasch model specifies a particular simple relationship between person ability, item difficulty, and the probability of observing a correct response. The implications of this specification are that:
>
> 1) the variable measured is unidimensional
> 2) there are no strong relationships among persons or items other than those specified by the model so that responses of persons to items are stochastically independent given their parameters in the model
> 3) items and persons do not differ substantially with respect to other possible response factors not represented in the model such as item discrimination, person sensitivity, guessing or indifference. (Wright and Mead, 1975, p. 2)

Thus, it will be necessary to analyze all of the response data collected using the MCC test. A computer program is available for these analyses.[1] Following analysis, if the test data provide person-free item calibration and item-free person measurement, then the three specifications of the model must have been met by the original design of the test.[2] These analyses will provide the calibration data needed for equating test difficulty levels and test content domains from grade 1 through college.

---

[1] CALFIT: Sample-free item calibration with a Rasch measurement model by Benjamin Wright and Ronald Mead, Statistical Laboratory, Department of Education, The University of Chicago, Chicago, Illinois, 1975. Note that this program is now operational at the State Education Department in Albany, New York.

[2] For details of the Rasch model, refer to Georg Rasch. Probabilistic models for some intelligence and attainment tests. Copenhagen: Nielsen & Lydiche, 1960; Benjamin Wright, Sample-free test calibration and person measurement, Proceedings of the 1967 Invitational Conference on Testing Problems, Princeton, N.J.: Educational Testing Service, 1968, pp.85-101.

Motivation for Using the Rasch Model[3]

Fifty years ago Thorndike complained that contemporary intelligence tests failed to specify "how far it is proper to add, subtract, multiply, divide, and compute ratios with the measures obtained" (Thorndike, 1926, p. 1). He asserted that a good measurement of ability would be one "on which zero will represent just not any of the ability in question, and 1, 2, 3, 4, and so on will represent amounts increasing by a constant difference " (Thorndike, 1926, p. 4).

Thorndike had the courage to complain because he believed he had worked out a solution to the problem for his own intelligence test. So did Thurstone (1925).

Thurstone's method was to translate the proportion in an age group passing any item into a unit normal deviate and to use these values as the basis for scaling. Common scale values for different age groups were obtained by assuming a linear relationship between the different scale values of overlapping items and using the different group means and standard deviations as the parameters for a transformation onto a common scale.

Thurstone redid a piece of Thorndike (actually Trabue's) work to show that the Thurstone method was superior (Thurstone, 1927). But the methods are essentially the same and they [illegible]re similar shortcomings.

Thurstone's "absolute scale" (1925, p. 438; 1927, pp. 518-19) yields an interval scale measurement of a kind. But no useful interpretation of the "equal" scale units has ever been proposed.

In addition to item homogeneity, the Thurstone method requires the assumption that ability is normally distributed with age groups and that there exist stable parameters for these distributions. Should the sampling of intended populations be biased, so will the scale values. They cannot be invariant to sampling. In particular, samples different in ability will produce scale values different in magnitude and dispersion.

---

[3] This section and the one that follows on "Application of the Rasch Model" were written by Dr. Benjamin Wright and Ronald Mead at the Department of Education, University of Chicago, Chicago, Illinois, 1975.

Thurstone used the 1925 version of his method for the rest of his scaling life (e.g., Thurstone, 1947), but the majority of test calibrators have relied on the simpler techniques of percentile ranks and standard scores. The inadequacies of these methods were clarified by Loevinger's 1947 analysis of the construction and evaluation of tests of ability (Loevinger, 1947, p. 42).

Loevinger showed that test homogeneity and scale monotonicity were essential criteria for adequate measurement. In addition, "An acceptable method of scaling must result in a derived scale which is independent of the original scale and of the original group tested " (Loevinger, 1947, p. 46).

Summing up the test calibration situation in 1947, Loevinger says "No system of scaling has been proved adequate by the criteria proposed here, though these criteria correspond to the claims made by Thurstone's system" (Loevinger, 1947, p. 43). As for reliabilities based on correlations, "Until an adequate system of scaling is found, the correlation between tests of abilities, even between two tests of the same ability, will be accidental to an unknown degree " (Loevinger, 1947, p. 46).

Twenty-five years ago Gulliksen concluded his Theory of Mental Tests (1950) with the following observation:

> Relatively little experimental or theoretical work has been done on the effect of group changes on item parameters. If we assume that a given item requires a certain ability (A), the proportion of a group answering that item correctly will increase and decrease as the ability level of the group changes. The amount of this change will be greater for an item that is highly correlated with ability A than for one that correlates only moderately with ability A.

> If we have some standard measure of ability A, it may be that the ability level at which 50 percent pass and 50 percent fail would not be subject to as much fluctuation as the proportion of correct responses. As yet there has been no systematic theoretical treatment of measures of item difficulty directed particularly toward determining the nature of their variation with respect to changes in group ability. Neither has the experimental work on item analysis been directed toward determining the relative invariance of item parameters with systematic changes in the ability level of the group tested. (Gulliksen, 1950, pp. 392-93)

At the 1953 E.T.S. Invitational Conference on Testing Problems, Tucker suggested that "An ideal test may be conceived as one for which the information transmitted by each of the possible scaled scores represents a location on some unitary continuum so that uniform differences between scaled scores correspond to uniform differences between test performances for all score levels " (Tucker, 1953, p. 27). He also proposed the comparison of groups differing in ability as a strong method for evaluating test homogeneity (Tucker, 1953, p. 25). But the other participants in the conference belittled his proposals as impractically idealistic.

Fifteen years ago Angoff wrote in an encyclopedia article on measurement and scaling:

> Most of the test scales now in use derive their systems of unit from data taken from actual test administrations, and thus are dependent on the performance of the groups tested. When so constructed, the scale has meaning only so long as the group is well defined and has meaning, and bears a resemblance in some fashion to the groups or individuals who later take the test for the particular purposes of selection, guidance, or group evaluation. However, if it is found that the sampling for the development of a test scale has been adequate, or that the group on which the test has been scaled has outlived its usefulness, possibly because of changes in the defined population or because of changes in educational emphases, then the scale itself comes into question. This is a serious matter. A test which is to have continued usefulness must have a scale which does not change with the times, which will permit acquaintance and familiarity with the system of units, and which will permit an accumulation of data for historical comparisons. (Angoff, 1960, p. 815)

And yet the faulted methods referred to and criticized by Loevinger, Gulliksen and Angoff are still widely used in test construction and measurement in spite of the fact that considerable evidence has accumulated in the past fifteen years that much better methods are available and practical.

The new attack on mental measurement was first formulated nearly twenty-five years ago by a Danish mathematician, Georg Rasch. Rasch began his work on psychological measurement in 1945 when he standardized a group intelligence test for the Danish Department of Defense. It was in carrying out that item analysis that he first "became aware of the problem of defining the difficulty of an item independently of the population and the ability of an individual independently of which items he has actually solved " (Rasch, 1960, viii). By 1952 he had laid down the basic foundations for his new psychometrics and worked out two probability models for the analysis of oral reading tests. In 1953 he reanalyzed the intelligence test data and developed the essentials of a probability model for item analysis.

Rasch first published his concern about the problem of sample dependent estimates in his 1953 article on simultaneous factor analysis in several populations (Rasch, 1953). But his work on item analysis was unknown in this country until the spring of 1960 when he visited Chicago for three months, gave a paper at the Berkeley Symposium on Mathematical Statistics (Rasch, 1961), and published a book, Probabilistic Model for Some Intelligence and Attainment Tests (Rasch, 1960).

These publications contain a detailed presentation and application of a probability model for the analysis of psychological test data (Rasch, 1960, pp. 73-79; pp. 107-125; pp. 168-182; 1961, pp. 323-325). The application of the model yields measurements which satisfy Loevinger's and Tucker's criteria and resolve the problems outlined by Gulliksen and Angoff. But unfortunately even after 1960 not many social scientists learned of Rasch's work. . . . Rasch's book, published in Denmark, reached only a handful of scholars in this country. His work has crucial implications for the future of psychometrics and for measurement in social science research in general.

Of the 1960 book, Tucker says "The monograph by Rasch presents several very interesting and quite sophisticated developments in mathematical test theory " (Tucker, 1963, p. 356). Of the item analysis model Sitgreaves says "the model proposed and the questions that it raises are extremely interesting. Over-all, the author has made a substantial contribution to model building in tests of ability " (Sitgreaves, 1963, p. 220). Coombs says that Rasch's work is a "major contribution and a new approach in psychometrics which is worthy of very serious study " (Coombs, 1964, p. 238).

In her discussion of person and population as psychometric concepts, Loevinger writes:

> Rasch (1960) has devised a truly new approach to psychometric problems ... He makes use of none of the classical psychometrics, but rather applies algebra anew to a probabilistic model. The probability that a person will answer an item correctly is assumed to be the product of an ability parameter pertaining only to the person and a difficulty parameter pertaining only to the item. Beyond specifying one person as the standard of ability and one item as the standard of difficulty, the ability assigned to an individual is independent of that of other members of the group and of the particular items with which he is tested; similarly for the item difficulty ... Indeed, these two properties were once suggested as criteria for absolute scaling (Loevinger, 1947); at that time proposed schemes for absolute scaling had not been shown to satisfy the criteria, nor does Guttman scaling do so. Thus, Rasch must be credited with an outstanding contribution to one of the two central psychometric problems, the achievement of nonarbitrary measures. Rasch is concerned with a different and more rigorous kind of generalization than Cronbach, Rajaratnam, and Gleser. When his model fits, the results are independent of the sample of persons and

> of the particular items within some broad limits. Within these limits, generality is, one might say, complete. (Loevinger, 1965, p. 151)

When a new method for solving old problems is proposed, an important question is, Does it work in practice?

Application of the Rasch Model

In his 1960 book, Rasch documents at length the application of his model to a four-test intelligence battery used by the Danish army (Rasch, 1960, pp. 80-107). The model is plainly inappropriate for two of the tests but a good fit to the other two. In subsequent, but still unreported analyses of these data, Rasch was able to track down a test administration factor causing one of these tests not to fit the model and to show that upon adjustment for this factor, the residual data of this test also fit.

Brooks and Blommers (1965) applied Rasch's model to Lorge-Thorndike Intelligence Tests administered to eighth and tenth graders. Their purpose was to evaluate the stability of test item parameters when given to groups of different ability. They found that the model fit the Lorge-Thorndike data rather well and that the estimates of item difficulty were stable.

Since then there have been a series of applications made in Denmark by Rasch's students. For example, Andersen (1964) applied a multiple-response generalization of the model to an attitude inventory administered to Danish recruits. He was able to show that his original test contained two homogeneous subsets of items which, when identified and isolated, each in themselves fit the model well. Andersen's subsequent work on the mathematical and statistical aspects of the model has been extensive (see Andersen's references).

In 1968, Wright applied the model to 48 reading comprehension items on the Law School Admission Test. He demonstrated the sample-freeness of the calibrations by estimating item difficulties separately for the highest and lowest scoring groups. Since the difficulty estimates based on the extreme groups were statistically equivalent, he had shown that the estimates were independent of the ability of the persons in the calibration sample and could be safely used over the entire range of ability. This method of demonstrating the practical utility of the model has been successfully applied on numerous ocassions (e.g., on the more than 50 different sets of test data brought by participants to the AERA Presessions on the Rasch model held in 1969, 1970, and 1975).

Durovic (1970) reported the successful application of the Rasch model to test development for the New York State Department of Civil Service. He found it especially useful for identifying poor items. In the several examples of aptitude and achievement type tests that he has investigated, items identified as misfitting were easily recognized subsequently as defective for clear-cut substantive reasons. The bad items typically required types of behavior or specific prior knowledge not needed for other items.

The American Guidance Service has used the model in their test construction work since 1970. Two of their tests, KEYMATH (Connally, Nachtman, and Pritchett, 1971) and the Woodcock Reading Mastery Test (Woodcock, 1974) were entirely built on Rasch principles. This involved not only the selection and calibration of items but also the development of recording forms which relate the tested person's estimated ability, in a criterion way, to the specific skills and deficiencies he has and, in a normative way, to his grade level.

Willmott and Fowles (1974) have also reported extensive application of the model in England. In connection with the Sixteen Plus Examining Project at the National Foundation for Educational Research of England and Wales, they applied it successfully to tests of reading ability, English comprehension, geography, science, mathematics and physics. While discussing questions of chaining items and building item pools, their emphasis was on fit to the model. They concluded that while to obtain the maximum benefits of the model it is necessary to take the trouble to construct a homogeneous set of items, this was not mandatory in order to use the model to obtain measurements far better than those ordinarily available.

Spada and Fischer (1973) used the framework of the logistic latent trait model for a scientific analysis of a projective inkblot test. They were able to formulate the specific (and conflicting) models of personality implied by the coding and scoring rules of the Rorschach and Holtzman tests. In an empirical test on 350 Rorschach and 305 Holtzman protocols, the Holtzman approach to scoring, which corresponds to that required by the Rasch model, was found to represent the data more adequately than the Rorschach scoring. The tests of fit were found useful in identifying misfitting inkblots and in modifying them to provoke more interpretable responses.

Bashaw (1974) and Rentz (1974) have completed a successful Rasch equating of seven reading tests used in the National Anchor Test Study by calibrating all items on all forms and all levels of each test on a common scale. Their results are essentially equivalent to the far more costly and awkward methods employed by ETS in the "official" equating but required half as much data, a third as much time and one tenth the processing budget. This demonstrated dramatically the simplicity and utility of the Rasch model over alternative methods of test equating.

Kifer and Bramble (1974) used a final examination constructed around performance objectives to illustrate the model's application to criterion-referenced testing. They discuss how to select items after the criterion is set and how to control the two types of classification errors. The standard errors of measurement that the model provides for each ability estimate make it possible to compute explicitly the probabilities that a person classified as a "master" actually lies below the criterion and that a person classified as a "non-master" actually lies above.

Reckase (1975) used simulated data to investigate the utility of the Rasch model in connection with tailored testing. His results indicated, that with reasonable stopping rules, the estimated abilities converged quickly to the true value (only eight or ten items were required in some cases). He also found that badly off-target tests produced biased estimates.

MCC Item Calibration Using the Rasch Model

The two previous sections in this report have supported the application of Rasch measurement models in educational test development. During the spring of 1975, thirty-six MCC test forms were administered to 5,000 urban students in grades [illegible] through 9. These MCC test forms were constructed from passages in the Reading/Literature section of the Textual Domain. All of these MCC test forms have been analyzed using the Rasch measurement model. A complete description of a Rasch analysis for one test form will provide a basis for understanding Rasch item/passage analysis, item/passage calibration, and test equating.

Figure 7.1 displays the distribution of subjects by total test scores on Form 14 of the MCC test. This histogram is scaled to fill the page; thus, the modal score (i.e., the measure that occurs most frequently in the distribution) is displayed as 100 percent. The modal value for Form 14 is a score of 51 out of a possible 60 indicating that the 4th, 5th, and 6th graders who took Form 14 did quite well on it.[4]

Figure 7.2 displays the number of subjects who answered each item correctly and the proportion of correct responses on each item. In Rasch terminology, this is a distribution of item easiness. Inspection of Figure 7.2 with horizontal lines drawn between items for the six different passages reveals the tendency for item easiness to change with the different passages.

Table 7.1 shows the results of the estimation process in the CALFIT computer program as developed by Wright and Mead (1975). The unconditional maximum likelihood procedure was used for these estimates of item difficulty

---

[4] When distributions are skewed like the distribution in Figure 7.1 for MCC Form 14, an unconditional maximum liklihood estimation routine is used in the CALFIT computer program for Rasch analyses. This routine provides more accurate estimates for skewed data than the less expensive, but approximate, estimation routine.

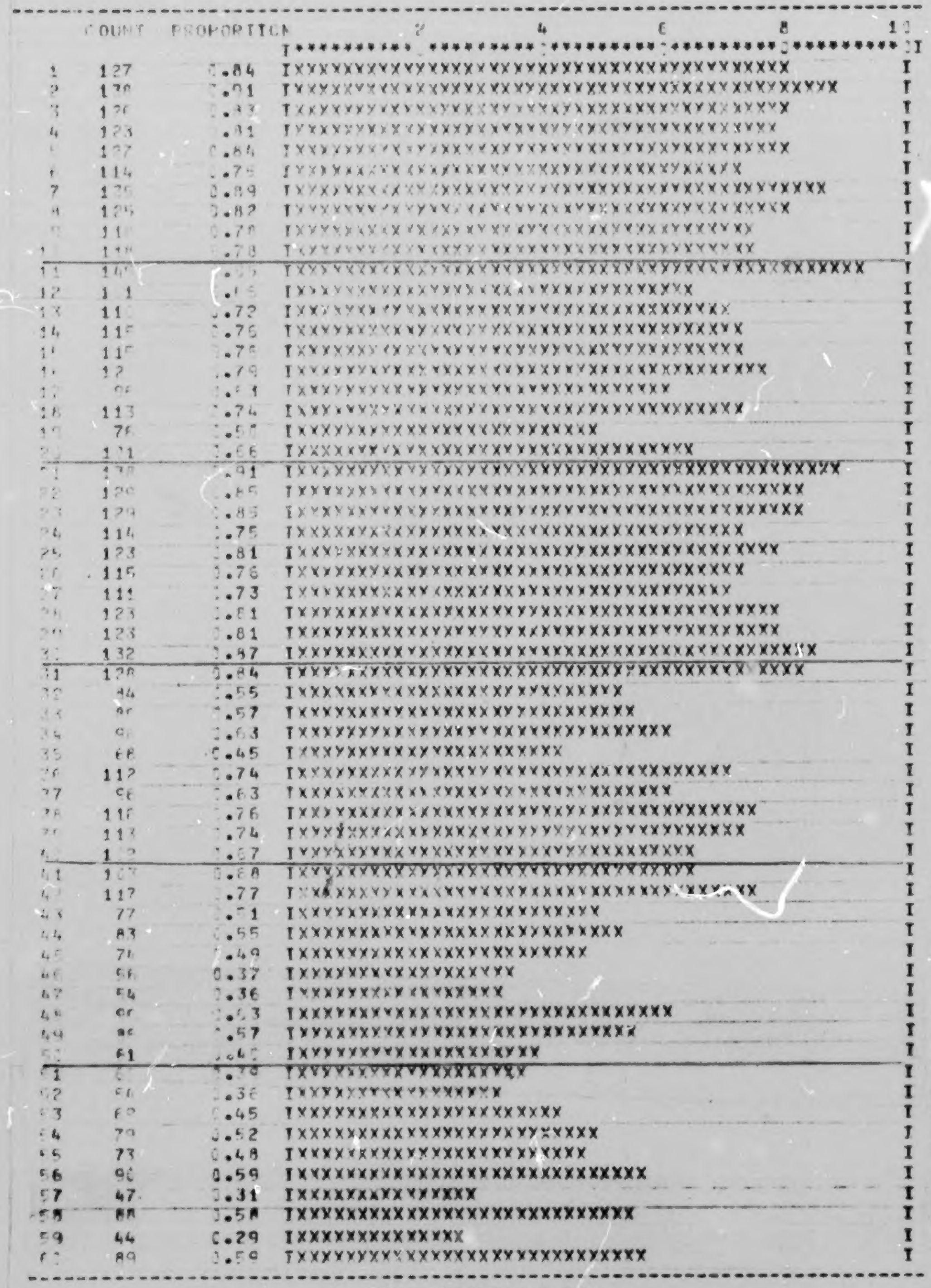
BEST COPY AVAILABLE

```
SCORE           DISTRIBUTION OF ABILITY
-------------------------------------------------------------------------------
        COUNT   PROPORTION          2          4          6          8          10
                          I**********0**********0**********0**********0**********0I
   1      0       0.00    I                                                       I
   2      1       0.01    IXXXXX                                                  I
   3      0       0.00    I                                                       I
   4      0       0.00    I                                                       I
   5      0       0.00    I                                                       I
   6      0       0.00    I                                                       I
   7      1       0.01    IXXXXX                                                  I
   8      1       0.01    IXXXXX                                                  I
   9      0       0.00    I                                                       I
  10      2       0.01    IXXXXXXXXXXX                                            I
  11      0       0.00    I                                                       I
  12      1       0.01    IXXXXX                                                  I
  13      4       0.03    IXXXXXXXXXXXXXXXXXXXXXX                                 I
  14      3       0.02    IXXXXXXXXXXXXXXXX                                       I
  15      1       0.01    IXXXXX                                                  I
  16      1       0.01    IXXXXX                                                  I
  17      1       0.01    IXXXXX                                                  I
  18      2       0.01    IXXXXXXXXXXX                                            I
  19      0       0.00    I                                                       I
  20      3       0.02    IXXXXXXXXXXXXXXXX                                       I
  21      1       0.01    IXXXXX                                                  I
  22      1       0.01    IXXXXX                                                  I
  23      0       0.00    I                                                       I
  24      1       0.01    IXXXXX                                                  I
  25      0       0.00    I                                                       I
  26      4       0.03    IXXXXXXXXXXXXXXXXXXXXXX                                 I
  27      5       0.03    IXXXXXXXXXXXXXXXXXXXXXXXXXXX                            I
  28      1       0.01    IXXXXX                                                  I
  29      4       0.03    IXXXXXXXXXXXXXXXXXXXXXX                                 I
  30      0       0.00    I                                                       I
  31      3       0.02    IXXXXXXXXXXXXXXXX                                       I
  32      1       0.01    IXXXXX                                                  I
  33      1       0.01    IXXXXX                                                  I
  34      3       0.02    IXXXXXXXXXXXXXXXX                                       I
  35      3       0.02    IXXXXXXXXXXXXXXXX                                       I
  36      2       0.01    IXXXXXXXXXXX                                            I
  37      3       0.02    IXXXXXXXXXXXXXXXX                                       I
  38      2       0.01    IXXXXXXXXXXX                                            I
  39      1       0.01    IXXXXX                                                  I
  40      5       0.03    IXXXXXXXXXXXXXXXXXXXXXXXXXXX                            I
  41      5       0.03    IXXXXXXXXXXXXXXXXXXXXXXXXXXX                            I
  42      7       0.05    IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX                 I
  43      4       0.03    IXXXXXXXXXXXXXXXXXXXXXX                                 I
  44      3       0.02    IXXXXXXXXXXXXXXXX                                       I
  45      5       0.03    IXXXXXXXXXXXXXXXXXXXXXXXXXXX                            I
  46      5       0.03    IXXXXXXXXXXXXXXXXXXXXXXXXXXX                            I
  47      6       0.04    IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX                      I
  48      3       0.02    IXXXXXXXXXXXXXXXX                                       I
  49      3       0.02    IXXXXXXXXXXXXXXXX                                       I
  50      5       0.03    IXXXXXXXXXXXXXXXXXXXXXXXXXXX                            I
  51      9       0.06    IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXI
  52      2       0.01    IXXXXXXXXXXX                                            I
  53      8       0.05    IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX           I
  54      5       0.03    IXXXXXXXXXXXXXXXXXXXXXXXXXXX                            I
  55      2       0.01    IXXXXXXXXXXX                                            I
  56      5       0.03    IXXXXXXXXXXXXXXXXXXXXXXXXXXX                            I
  57      4       0.03    IXXXXXXXXXXXXXXXXXXXXXX                                 I
  58      6       0.04    IXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX                      I
  59      3       0.02    IXXXXXXXXXXXXXXXX                                       I
  60      0       0.00    I                                                       I
-------------------------------------------------------------------------------
                            FULL SCALE = 0.06
```

Figure 7.1. Rasch analysis of MCC Form 14: distribution of subjects by total test score.

BEST COPY AVAILABLE

ITEM DISTRIBUTION OF EASINESS



ITEM	COUNT	PROPORTION
1	127	0.84
2	138	0.91
3	126	0.83
4	123	0.81
5	127	0.84
6	114	0.75
7	135	0.89
8	125	0.82
9	118	0.78
10	118	0.78
11	14[illegible]	0.[illegible]
12	101	0.66
13	109	0.72
14	115	0.76
15	115	0.76
16	120	0.79
17	93	0.61
18	113	0.74
19	76	0.50
20	101	0.66
21	138	0.91
22	129	0.85
23	129	0.85
24	114	0.75
25	123	0.81
26	115	0.76
27	111	0.73
28	123	0.81
29	123	0.81
30	132	0.87
31	128	0.84
32	84	0.55
33	86	0.57
34	96	0.63
35	68	0.45
36	112	0.74
37	96	0.63
38	116	0.76
39	113	0.74
40	102	0.67
41	103	0.68
42	117	0.77
43	77	0.51
44	83	0.55
45	74	0.49
46	56	0.37
47	54	0.36
48	96	0.63
49	86	0.57
50	61	0.40
51	58	0.38
52	56	0.36
53	68	0.45
54	79	0.52
55	73	0.48
56	90	0.59
57	47	0.31
58	88	0.58
59	44	0.29
60	89	0.59

FULL SCALE = 1.00

Figure 7.2. Rasch analysis of MCC Form 14: distribution of easiness--subjects answering items correctly.

Table 7.1

Rasch Analysis of MCC Form 14:
Estimation of Item Difficulty and Group Ability

```
PROCEDURE USED UCON
NUMBER OF ITERATIONS =    3

------------------------------------------------------------------------------------------------
SEQUENCE I   ITEM I     ITEM       STANDARD    LAST DIFF I  SCORE     GROUP     STANDARD
 NUMBER  I   NAME I  DIFFICULTY     ERROR       CHANGE   I  GROUP    ABILITY      ERROR
------------------------------------------------------------------------------------------------
    1    I   P1I1 I    -1.188       0.257       -0.019   I     1      -4.71        1.03
    2    I   P1I2 I    -2.062       0.316       -0.031   I     2      -3.97        0.74
    3    I   P1I3 I    -1.123       0.254       -0.018   I     3      -3.52        0.62
    4    I   P1I4 I    -0.938       0.245       -0.015   I     4      -3.19        0.54
    5    I   P1I5 I    -1.188       0.257       -0.019   I     5      -2.92        0.49
    6    I   P1I6 I    -0.448       0.226       -0.007   I     6      -2.69        0.46
    7    I   P1I7 I    -1.784       0.294       -0.028   I     7      -2.50        0.43
    8    I   P1I8 I    -1.060       0.251       -0.017   I     8      -2.32        0.41
    9    I   P1I9 I    -0.656       0.233       -0.011   I     9      -2.16        0.39
   10    I  P1I10 I    -0.656       0.233       -0.011   I    10      -2.02        0.38
   11    I   P2I1 I    -2.971       0.418       -0.040   I    11      -1.88        0.36
   12    I   P2I2 I     0.154       0.209        0.003   I    12      -1.75        0.35
   13    I   P2I3 I    -0.253       0.220       -0.004   I    13      -1.63        0.34
   14    I   P2I4 I    -0.499       0.228       -0.008   I    14      -1.51        0.34
   15    I   P2I5 I    -0.499       0.228       -0.008   I    15      -1.40        0.33
   16    I   P2I6 I    -0.766       0.238       -0.012   I    16      -1.30        0.32
   17    I   P2I7 I     0.365       0.205        0.006   I    17      -1.19        0.32
   18    I   P2I8 I    -0.398       0.224       -0.006   I    18      -1.09        0.31
   19    I   P2I9 I     1.151       0.197        0.018   I    19      -1.00        0.31
   20    I  P2I10 I     0.154       0.209        0.003   I    20      -0.90        0.31
   21    I   P3I1 I    -2.062       0.316       -0.031   I    21      -0.81        0.30
   22    I   P3I2 I    -1.323       0.264       -0.021   I    22      -0.72        0.30
   23    I   P3I3 I    -1.323       0.264       -0.021   I    23      -0.63        0.30
   24    I   P3I4 I    -0.448       0.226       -0.007   I    24      -0.54        0.30
   25    I   P3I5 I    -0.938       0.245       -0.015   I    25      -0.45        0.30
   26    I   P3I6 I    -0.499       0.228       -0.008   I    26      -0.36        0.30
   27    I   P3I7 I    -0.300       0.221       -0.005   I    27      -0.27        0.30
   28    I   P3I8 I    -0.938       0.245       -0.015   I    28      -0.19        0.29
   29    I   P3I9 I    -0.938       0.245       -0.015   I    29      -0.10        0.29
   30    I  P3I10 I    -1.541       0.277       -0.024   I    30      -0.01        0.29
   31    I   P4I1 I    -1.254       0.260       -0.020   I    31       0.07        0.29
   32    I   P4I2 I     0.944       0.199        0.014   I    32       0.16        0.30
   33    I   P4I3 I     0.766       0.200        0.013   I    33       0.25        0.30
   34    I   P4I4 I     0.365       0.205        0.006   I    34       0.34        0.30
   35    I   P4I5 I     1.456       0.197        0.023   I    35       0.43        0.30
   36    I   P4I6 I    -0.349       0.223       -0.005   I    36       0.52        0.30
   37    I   P4I7 I     0.365       0.205        0.006   I    37       0.61        0.30
   38    I   P4I8 I    -0.550       0.229       -0.009   I    38       0.70        0.31
   39    I   P4I9 I    -0.398       0.224       -0.006   I    39       0.79        0.31
   40    I  P4I10 I     0.110       0.210        0.002   I    40       0.89        0.31
   41    I   P5I1 I     0.067       0.211        0.001   I    41       0.99        0.31
   42    I   P5I2 I    -0.603       0.231       -0.010   I    42       1.09        0.32
   43    I   P5I3 I     1.113       0.197        0.018   I    43       1.19        0.32
   44    I   P5I4 I     0.882       0.198        0.015   I    44       1.30        0.33
   45    I   P5I5 I     1.227       0.197        0.020   I    45       1.41        0.33
   46    I   P5I6 I     1.924       0.201        0.029   I    46       1.52        0.34
   47    I   P5I7 I     2.004       0.202        0.030   I    47       1.64        0.35
   48    I   P5I8 I     0.365       0.205        0.006   I    48       1.76        0.36
   49    I   P5I9 I     0.766       0.200        0.013   I    49       1.89        0.37
   50    I  P5I10 I     1.726       0.199        0.026   I    50       2.03        0.38
-------------------------------------------------------------------------------------------------
   51    I   P6I1 I     1.766       0.199        0.027   I    51       2.18        0.39
   52    I   P6I2 I     2.004       0.202        0.030   I    52       2.34        0.41
   53    I   P6I3 I     1.418       0.197        0.022   I    53       2.52        0.43
   54    I   P6I4 I     1.036       0.197        0.017   I    54       2.72        0.46
   55    I   P6I5 I     1.265       0.197        0.020   I    55       2.94        0.49
   56    I   P6I6 I     0.608       0.201        0.010   I    56       3.21        0.54
   57    I   P6I7 I     2.295       0.208        0.033   I    57       3.54        0.61
   58    I   P6I8 I     0.687       0.200        0.011   I    58       3.99        0.74
   59    I   P6I9 I     2.425       0.211        0.035   I    59       4.72        1.02
   60    I  P6I10 I     0.648       0.201        0.011   I
------------------------------------------------------------------------------------------------
                  ROOT MEAN SQUARE =     0.018
```

and group ability. "The information in the table is organized in two sections. The left side reports the item estimation process. For each item, its difficulty estimate and the standard error of this estimate are given. Items are identified by sequence number, which is internal to a given run and would change if items were added or deleted in other runs, and by item name, a four-character alphameric, supplied by the user.[5] . . .The right side of the table contains the relation between observable test score and the corresponding estimates of ability. The 'ability' column contains the estimate of ability implied by each possible score " (Wright and Mead, 1975, pp. 16-17). The "Score Group" and "Group Ability" columns in Table 7.1 indicate that a student who obtained a score of 45 on these Form 14 items, as calibrated, would be assigned an ability of 1.41, with a standard error of 0.33 for that estimated ability.

Table 7.2 begins an analysis of the fit of the data to the Rasch model. For these test data, the students were separated into six groups, by score. "The table contains one row for each item, identified both by internal sequence number and user supplied item name. . . . The body of the table is in three sections: The left six columns contain the item characteristic curves, the proportion of subjects in each group who answered each item correctly. This should approximate the shape of the logistic curve for items that fit the model. . . . The center six columns contain the number of answers unexplained by the model. It is computed as the number of correct answers observed in a group minus the number that would be predicted by the

---

[5] The user name in the Rasch analysis of the MCC Test indicates an item in a particular passage on the test form, e.g., P1I1 = passage 1, item 1; P2I4 = passage 2, item 4; and so forth.

Table 7.2

Rasch Fit Analysis of MCC Form 14

			ITEM CHARACTERISTIC CURVE							NUMBER UNEXPECTED ANSWERS							FIT Z-SQUARED							
SEQ NUM	ITEM NAME	I	1ST GROUP	2ND GROUP	3RD GROUP	4TH GROUP	5TH GROUP	6TH GROUP	I	1ST GROUP	2ND GROUP	3RD GROUP	4TH GROUP	5TH GROUP	6TH GROUP	I	1ST GRP	2ND GRP	3RD GRP	4TH GRP	5TH GRP	6TH GRP	I	FIT MN.SQ.
1	P1I1	I	0.75	0.76	0.84	0.88	0.78	1.00	I	7.7	0.0	-1.2	-1.3	-5.1	0.3	I	13.0	0.0	0.7	1.3	38.6	0.3	I	8.98
2	P1I2	I	0.47	0.60	1.00	1.00	0.96	1.00	I	0.9	-2.1	1.3	0.7	-0.6	0.1	I	0.2	2.0	1.0	0.9	1.3	0.1	I	1.04
3	P1I3	I	0.42	0.80	0.84	0.92	1.00	0.96	I	0.0	1.3	-1.0	-0.2	0.9	-0.7	I	0.0	0.5	0.0	0.0	1.2	2.2	I	0.72
4	P1I4	I	0.37	0.68	0.84	0.96	0.96	1.00	I	-0.0	-0.8	-0.5	1.2	0.1	0.3	I	0.0	0.1	0.1	0.8	0.0	0.4	I	0.25
5	P1I5	I	0.42	0.64	1.00	1.00	0.96	0.96	I	-0.3	-3.0	2.8	1.7	-0.1	-0.7	I	0.0	2.4	3.0	2.3	0.0	2.5	I	1.85
6	P1I6	I	0.37	0.56	0.80	0.92	0.99	0.92	I	2.4	-1.1	0.3	1.3	-1.2	-1.4	I	1.5	0.2	0.0	0.8	1.2	4.7	I	1.41
7	P1I7	I	0.50	0.84	1.00	1.00	0.96	1.00	I	-1.6	-0.3	1.7	1.0	-0.5	0.1	I	0.6	0.0	2.2	1.2	0.7	0.2	I	0.81
8	P1I8	I	0.37	0.76	0.92	0.96	0.96	0.92	I	-0.7	0.6	1.2	0.9	-0.0	-1.7	I	0.1	0.1	0.6	0.6	0.0	11.7	I	2.18
9	P1I9	I	0.33	0.80	0.72	0.85	0.96	0.96	I	0.4	3.7	-2.5	-1.2	0.4	-0.5	I	0.0	3.0	2.1	0.7	0.2	0.8	I	1.15
10	P1I10	I	0.25	0.64	0.92	0.92	0.96	0.92	I	-1.6	-0.3	2.5	0.8	0.4	-1.5	I	0.6	0.0	2.0	0.3	0.2	6.6	I	1.63
11	P2I1	I	0.83	0.92	0.96	1.00	1.00	1.00	I	1.0	-0.7	-0.5	0.3	0.2	0.0	I	0.3	0.5	0.5	0.4	0.2	0.1	I	0.33
12	P2I2	I	0.29	0.40	0.72	0.73	0.89	0.92	I	2.8	-1.4	1.2	-1.5	0.0	-1.0	I	2.7	0.4	0.3	0.6	0.0	1.3	I	0.90
13	P2I3	I	0.37	0.52	0.56	0.88	0.96	1.00	I	3.2	-0.9	-4.9	1.0	1.1	0.7	I	3.0	0.2	6.3	0.3	0.8	0.8	I	1.91
14	P2I4	I	0.33	0.60	0.84	0.85	0.89	1.00	I	1.2	-0.4	1.1	-0.8	-1.3	0.5	I	0.3	0.0	0.3	0.3	1.4	0.7	I	0.51
15	P2I5	I	0.25	0.60	0.84	0.92	0.89	1.00	I	-0.8	-0.4	1.1	1.2	-1.3	0.5	I	0.2	0.0	0.3	0.6	1.4	0.7	I	0.54
16	P2I6	I	0.29	0.72	0.84	0.88	0.96	1.00	I	-1.1	1.1	0.1	-0.5	0.3	0.4	I	0.3	0.3	0.	0.1	0.1	0.5	I	0.22
17	P2I7	I	0.17	0.32	0.56	0.77	0.96	0.96	I	0.4	-2.2	-1.6	0.5	2.7	0.2	I	0.1	1.0	0.6	0.1	2.8	0.0	I	0.75
18	P2I8	I	0.25	0.56	0.92	0.77	0.93	1.00	I	-1.4	-0.8	3.5	-2.5	-0.2	0.6	I	0.0	0.1	3.5	2.6	0.0	0.7	I	1.16
19	P2I9	I	0.21	0.20	0.64	0.58	0.67	0.68	I	3.2	-1.0	5.2	-0.1	-2.1	-5.5	I	7.5	0.3	5.3	0.0	1.1	17.2	I	5.24
20	P2I10	I	0.25	0.44	0.64	0.81	0.85	0.96	I	1.8	-0.4	-0.8	0.5	-1.0	-0.0	I	1.1	0.0	0.2	0.1	0.4	0.0	I	0.31
21	P3I1	I	0.54	0.92	1.00	0.96	1.00	1.00	I	-2.1	0.3	1.3	-0.3	0.4	0.1	I	1.0	0.4	1.0	0.1	0.5	0.1	I	0.63
22	P3I2	I	0.42	0.92	0.96	0.85	0.93	1.00	I	-1.1	3.4	1.5	-2.5	-1.2	0.2	I	0.3	3.4	1.2	5.3	2.5	0.3	I	2.13
23	P3I3	I	0.33	0.88	0.96	1.00	1.00	0.96	I	-3.1	0.4	1.5	1.5	0.8	-0.8	I	2.1	0.1	1.2	2.0	1.0	3.1	I	1.57
24	P3I4	I	0.33	0.60	0.84	0.85	0.85	1.00	I	1.4	-0.1	1.3	-0.7	-2.2	0.6	I	0.5	0.0	0.5	0.2	3.8	0.7	I	0.94
25	P3I5	I	0.25	0.68	0.92	0.96	1.00	1.00	I	-3.0	-0.8	1.5	1.2	1.1	0.3	I	2.1	0.1	[illegible]	0.8	1.4	0.4	I	0.98
26	P3I6	I	0.21	0.64	0.92	0.81	0.96	0.96	I	-1.8	0.6	3.1	-1.8	0.7	-0.5	I	0.9	0.1	2.0	1.4	0.4	0.5	I	1.02
27	P3I7	I	0.08	0.68	0.68	0.88	1.00	1.00	I	-4.0	2.8	-2.1	0.8	2.0	0.6	I	4.5	1.6	1.2	0.2	2.7	0.8	I	1.83
28	P3I8	I	0.17	0.76	0.96	0.92	1.00	1.00	I	-5.0	1.2	2.5	0.2	1.1	0.3	I	5.8	0.4	2.5	0.0	1.4	0.4	I	1.77
29	P3I9	I	0.25	0.68	0.96	0.96	0.96	1.00	I	-3.0	-0.8	2.5	1.2	0.1	0.3	I	2.1	0.1	2.0	0.8	0.0	0.4	I	1.12
30	P3I10	I	0.37	0.84	0.96	1.00	1.00	1.00	I	-3.3	0.6	1.1	1.2	0.6	0.2	I	2.4	0.1	0.7	1.6	0.8	0.2	I	0.97
31	P4I1	I	0.29	0.84	0.92	1.00	1.00	0.96	I	-3.7	1.7	0.7	1.6	0.8	-0.7	I	3.0	0.8	0.2	2.1	1.0	2.8	I	1.67
32	P4I2	I	0.08	0.16	0.60	0.73	0.78	0.92	I	-0.4	-3.5	2.3	2.0	-0.6	-0.1	I	0.1	2.9	1.0	0.9	0.1	0.0	I	0.83
33	P4I3	I	0.04	0.32	0.48	0.69	0.85	0.96	I	-1.5	0.1	-1.2	0.6	1.1	0.7	I	1.3	0.0	0.3	0.1	0.4	0.4	I	0.41
34	P4I4	I	0.21	0.48	0.60	0.73	0.78	0.96	I	1.4	1.8	-0.6	-0.5	-2.3	0.2	I	0.8	0.7	0.1	0.1	2.2	0.0	I	0.64
35	P4I5	I	0.17	0.48	0.40	0.38	0.59	0.64	I	2.6	7.3	1.0	-3.1	-2.5	-5.8	I	6.6	16.9	0.2	1.9	1.3	15.1	I	6.90
36	P4I6	I	0.13	0.52	0.84	0.92	0.96	1.00	I	-3.2	-1.5	1.7	1.6	0.9	0.6	I	2.8	0.5	0.8	1.1	0.6	0.8	I	1.08
37	P4I7	I	0.17	0.36	0.68	0.73	0.89	0.92	I	0.4	-1.2	1.4	-0.5	0.7	-0.8	I	0.1	0.3	0.4	0.1	0.2	0.7	I	0.28
38	P4I8	I	0.17	0.56	0.92	0.92	0.96	1.00	I	-3.1	-1.7	2.9	1.1	0.6	0.5	I	2.5	0.6	2.6	0.5	0.3	0.6	I	1.18
39	P4I9	I	0.20	0.40	0.80	1.00	0.89	0.96	I	0.6	-2.8	0.5	3.5	-1.2	-0.4	I	0.1	1.6	0.1	5.0	1.0	0.4	I	1.35
40	P4I10	I	0.13	0.44	0.60	0.92	0.93	0.96	I	-1.4	-0.7	-2.1	3.3	0.9	-0.1	I	0.7	0.1	1.0	3.2	0.4	0.0	I	0.90

BEST COPY AVAILABLE

Table 7.2 (Continued)

```
                    ITEM CHARACTERISTIC CURVE                       NUMBER UNEXPECTED ANSWERS                              FIT Z-SQUARED
-------------------------------------------------------------------------------------------------------------------------------------------------------
SEQ  ITEM  I  1ST   2ND   3RD   4TH   5TH   6TH  I  1ST   2ND   3RD   4TH   5TH   6TH  I  1ST  2ND  3RD  4TH  5TH  6TH  I   FIT
NUM  NAME  I GROUP GROUP GROUP GROUP GROUP GROUP I GROUP GROUP GROUP GROUP GROUP GROUP I  GRP  GRP  GRP  GRP  GRP  GRP  I MN.SQ.
-------------------------------------------------------------------------------------------------------------------------------------------------------
 41  P5I1  I  0.17  0.40  0.68  0.77  1.00  1.00 I  -0.5  -2.0  -0.3  -0.9   2.8   0.9 I  0.1  0.8  0.0  0.2  3.9  1.2  I  1.03
 42  P5I2  I  0.21  0.60  0.80  0.96  1.00  1.00 I  -2.3  -1.0  -0.3   1.9   1.5   0.5 I  1.4  0.2  0.0  1.8  2.0  0.6  I  1.00
 43  P5I3  I  0.08  0.28  0.52  0.62  0.63  0.88 I   0.1   0.8   1.9   0.7  -3.3  -0.6 I  0.0  0.2  0.7  0.1  2.7  0.2  I  0.66
 44  P5I4  I  0.21  0.44  0.60  0.73  0.52  0.76 I   2.7   3.7   2.5   2.3  -7.4  -4.1 I  4.3  3.3  1.2  1.1 15.3 11.4  I  6.11
 45  P5I5  I  0.21  0.40  0.28  0.42  0.67  0.92 I   3.3   4.3  -3.4  -3.6  -1.7   0.6 I  8.6  5.3  2.3  2.5  0.7  0.2  I  3.26
 46  P5I6  I  0.17  0.12  0.20  0.23  0.63  0.84 I   3.1  -0.2  -1.6  -4.2   1.4   0.7 I 14.0  0.0  0.0  3.5  0.4  0.2  I  3.11
 47  P5I7  I  0.17  0.20  0.20  0.23  0.44  0.68 I   3.2   2.0  -1.2  -3.7  -3.0   2.0 I 16.7  1.9  0.4  2.7  1.7  1.3  I  3.95
 48  P5I8  I  0.08  0.48  0.44  0.77  0.96  1.00 I  -1.6   1.8  -4.6   0.5   2.7   1.2 I  1.0  0.7  4.5  0.1  2.8  1.6  I  1.77
 49  P5I9  I  0.17  0.28  0.32  0.62  1.00  0.96 I   1.5  -0.9  -5.2  -1.4   5.1   0.7 I  1.1  0.2  5.4  0.4  7.9  0.4  I  2.57
 50  P510  I  0.13  0.16  0.32  0.31  0.59  0.68 I   1.9   0.2   0.4  -3.4  -0.8   1.0 I  4.5  0.0  0.0  2.2  0.1  0.4  I  1.22
 51  P6I1  I  0.29  0.16  0.08  [illegible]  0.56  0.92 I   6.0   0.3  -5.3  -2.1  -1.6   2.1 I 44.6  0.0  6.8  0.9  0.5  1.7  I  9.08
 52  P6I2  I  0.[illegible]  0.16  0.28  0.42  0.44  0.72 I   1.2   1.[illegible]   0.8   1.3  -3.0  -2.0 I  2.2  0.5  0.2  0.4  1.7  1.3  I  1.04
 53  P6I3  I  0.[illegible]  0.2[illegible]  0.28  0.46  0.78  0.96 I  -1.4   0.1  -2.3  -1.4   2.3   2.1 I  1.8  0.0  1.1  0.4  1.2  2.1  I  1.08
 54  P6I4  I  0.08  0.32  0.24  0.54  0.89  1.00 I  -0.0   1.6  -5.5  -1.8   3.3   2.2 I  0.0  0.5  6.1  0.6  2.8  3.1  I  2.19
 55  P6I5  I  0.13  0.08  0.40  0.58  0.75  0.96 I   1.4  -3.5  -0.2   0.6  -0.5   1.7 I  1.5  3.5  0.[illegible]  0.1  0.1  1.5  I  1.12
 56  P6I6  I  0.[illegible]  0.20  0.52  0.81  0.96  1.00 I  -2.9  -3.8  -1.2   2.7   3.5   1.5 I  4.1  3.1  0.3  1.7  4.1  2.0  I  2.54
 57  P6I7  I  0.08  0.12  0.08  0.35  0.48  0.72 I   1.4   0.7  -2.9   1.0  -0.1  -0.8 I  3.9  0.3  2.7  0.2  0.0  0.2  I  1.21
 58  P6I8  I  0.[illegible]  0.24  0.48  0.77  0.93  1.00 I  -2.7  -2.3  -1.7   2.2   2.8   1.6 I  3.8  1.2  0.6  1.0  2.5  2.2  I  1.88
 59  P6I9  I  0.00  0.12  0.20  0.27  0.37  0.76 I  -0.5  [illegible]   0.6  -0.3  -2.2   0.8 I  0.7  0.6  0.1  0.0  0.9  0.1  I  0.41
 60  P610  I  0.17  0.24  0.56  0.54  0.56  1.00 I   1.2  -2.5   0.1  -4.1   3.6   1.6 I  0.7  1.4  0.[illegible]  3.7  4.3  2.1  I  2.03
-------------------------------------------------------------------------------------------------------------------------------------------------------
SCORE RANGE   1-25 26-35 36-42 43-48 49-53 54-59  N=  24    25    25    26    27    25      3.1  1.1  1.4  1.1  2.1  1.9   MEAN
MEAN ABILITY -1.52 -0.[illegible]2  0.88  1.48  2.23  3.51        PLUS=TOO MANY RIGHT                6.4  2.4  1.7  1.2  5.3  3.5   S.D.
                                                  MINUS=TOO MANY WRONG
```

7-21

BEST COPY AVAILABLE

model for a group of that size and ability on an item of that difficulty. A positive number indicates that the group did better on the item than we would expect from their performance on the other items. A negative number indicates that they did worse. . . . The third section of the table contains z^2 statistics for testing the fit of each item in each score group. They are approximately distributed as chi-square statistics with one degree of freedom. The first column on the right, 'FIT MN SQ,' contains a statistic for testing the fit of each item over all groups. Since the deviations from the model were standardized in computing the z^2 statistics, the mean squares have expected values of one, and can be evaluated as F-ratios with numerator degrees of freedom equal to the number of groups minus one and infinite denominator degrees of freedom. . . . In addition to group definitions by score, the summary below the table includes the mean ability of each group and a count of the number of subjects in each group. The data below the z^2 table are the mean and standard deviation of the entries in each column. Under the hypothesis that the model fits the data, these have expected values of 1.0 for the means and 1.4 for the standard deviations" (Wright and Mead, 1975, pp. 17-18).

Table 7.3 contains fit information in three sequences: serial order, difficulty order, and fit order. "In each case the information given is item difficulty, an index of item discriminating power and the item fit mean square. . ." (Wright and Mead, 1975, p. 18). The item/total point biserial correlation is on the extreme right, in fit order.

"Under the left side of the table, the mean and standard deviation are given for item difficulty, discrimination, and fit mean square. Difficulty estimates are centered at zero and discrimination indices at one. The fit mean square has an expected mean of one and an expected standard deviation

Table 7.3

Rasch Fit Information on MCC Form 14 in Three Sequences--Serial, Difficulty, and Fit Orders

SERIAL ORDER						DIFFICULTY ORDER						FIT ORDER						
SEQ NUM	ITEM NAME	ITEM DIFF	DISC INDX	FIT MN SQ	I I	SEQ NUM	ITEM NAME	ITEM DIFF	DISC INDX	FIT MN SQ	I I	SEQ NUM	ITEM NAME	ITEM DIFF	DISC INDX	FIT MN SQ	I I	POINT BISER
1	P1I1	-1.19	-13.94	8.98	I	11	P2I1	-2.97	0.92	0.33	I	16	P2I6	-0.77	1.04	0.22	I	0.59
2	P1I2	-2.06	0.97	1.04	I	2	P1I2	-2.06	0.97	1.04	I	4	P1I4	-0.94	1.04	0.25	I	0.54
3	P1I3	-1.12	0.95	0.72	I	21	P3I1	-2.06	1.10	0.62	I	37	P4I7	0.36	0.98	0.27	I	0.56
4	P1I4	-0.94	1.04	0.25	I	7	P1I7	-1.78	1.10	0.80	I	20	P210	0.15	0.90	0.29	I	0.53
5	P1I5	-1.19	1.06	1.84	I	30	P310	-1.54	1.18	0.97	I	11	P2I1	-2.97	0.92	0.33	I	0.23
6	P1I6	-0.45	0.85	1.40	I	23	P3I3	-1.32	1.14	1.56	I	59	P6I9	2.42	0.95	0.40	I	0.45
7	P1I7	-1.78	1.10	0.80	I	22	P3I2	-1.32	0.90	2.17	I	33	P4I3	0.77	1.27	0.41	I	0.61
8	P1I8	-1.06	0.99	2.18	I	31	P4I1	-1.25	1.16	1.67	I	14	P2I4	-0.50	0.90	0.50	I	0.50
9	P1I9	-0.66	0.87	1.15	I	5	P1I5	-1.19	1.06	1.84	I	15	P2I5	-0.53	1.04	0.53	I	0.59
10	P110	-0.66	1.07	1.62	I	1	P1I1	-1.19	-13.94	8.98	I	21	P3I1	-2.06	1.10	0.62	I	0.51
11	P2I1	-2.97	0.92	0.33	I	3	P1I3	-1.12	0.95	0.72	I	34	P4I4	0.36	0.82	0.64	I	0.46
12	P2I2	0.15	0.83	0.90	I	8	P1I8	-1.06	0.99	2.18	I	43	P5I3	1.11	0.84	0.65	I	0.51
13	P2I3	-0.25	0.92	1.91	I	28	P3I8	-0.94	1.31	1.77	I	3	P1I3	-1.12	0.95	0.72	I	0.48
14	P2I4	-0.50	0.90	0.50	I	29	P3I9	-0.94	1.21	1.01	I	17	P2I7	0.36	1.19	0.74	I	0.61
15	P2I5	-0.53	1.04	0.53	I	4	P1I4	-0.94	1.04	0.25	I	7	P1I7	-1.78	1.10	0.80	I	0.50
16	P2I6	-0.77	1.04	0.22	I	25	P3I5	-0.94	1.22	0.98	I	32	P4I2	0.84	1.14	0.82	I	0.62
17	P2I7	0.36	1.19	0.74	I	16	P2I6	-0.77	1.04	0.22	I	40	P410	0.11	1.20	0.89	I	0.61
18	P2I8	-0.40	1.00	1.15	I	9	P1I9	-0.66	0.87	1.15	I	12	P2I2	0.15	0.83	0.90	I	0.48
19	P2I9	1.15	0.59	5.23	I	10	P110	-0.66	1.07	1.62	I	24	P3I4	-0.45	0.87	0.94	I	0.52
20	P210	0.15	0.90	0.29	I	42	P5I2	-0.60	1.22	0.99	I	30	P310	-1.54	1.18	0.97	I	0.61
21	P3I1	-2.06	1.10	0.62	I	38	P4I8	-0.55	1.28	1.17	I	25	P3I5	-0.94	1.22	0.98	I	0.64
22	P3I2	-1.32	0.90	2.17	I	15	P2I5	-0.53	1.04	0.53	I	42	P5I2	-0.60	1.22	0.99	I	0.68
23	P3I3	-1.32	1.14	1.56	I	14	P2I4	-0.50	0.90	0.50	I	29	P3I9	-0.94	1.21	1.01	I	0.67
24	P3I4	-0.45	0.87	0.94	I	26	P3I6	-0.50	1.07	1.02	I	26	P3I6	-0.50	1.07	1.02	I	0.58
25	P3I5	-0.94	1.22	0.98	I	6	P1I6	-0.45	0.85	1.40	I	41	P5I1	0.07	1.11	1.02	I	0.65
26	P3I6	-0.50	1.07	1.02	I	24	P3I4	-0.45	0.87	0.94	I	52	P6I2	2.00	0.74	1.03	I	0.41
27	P3I7	-0.30	1.32	1.83	I	39	P4I9	-0.40	0.99	1.34	I	2	P1I2	-2.06	0.97	1.04	I	0.43
28	P3I8	-0.94	1.31	1.77	I	18	P2I8	-0.40	1.00	1.15	I	36	P4I6	-0.35	1.34	1.08	I	0.70
29	P3I9	-0.94	1.21	1.01	I	36	P4I6	-0.35	1.34	1.08	I	53	P6I3	1.42	1.38	1.08	I	0.63
30	P310	-1.54	1.18	0.97	I	27	P3I7	-0.30	1.32	1.83	I	55	P6I5	1.27	1.24	1.12	I	0.57
31	P4I1	-1.25	1.16	1.67	I	13	P2I3	-0.25	0.92	1.91	I	9	P1I9	-0.66	0.87	1.15	I	0.47
32	P4I2	0.84	1.14	0.82	I	41	P5I1	0.07	1.11	1.02	I	18	P2I8	-0.40	1.00	1.15	I	0.59
33	P4I3	0.77	1.27	0.41	I	40	P410	0.11	1.20	0.89	I	38	P4I8	-0.55	1.28	1.17	I	0.67
34	P4I4	0.36	0.82	0.64	I	20	P210	0.15	0.90	0.29	I	57	P6I7	2.29	0.92	1.20	I	0.41
35	P4I5	1.46	0.38	6.98	I	12	P2I2	0.15	0.83	0.90	I	50	P510	1.73	0.92	1.22	I	0.46
36	P4I6	-0.35	1.34	1.08	I	37	P4I7	0.36	0.98	0.27	I	39	P4I9	-0.40	0.99	1.34	I	0.60
37	P4I7	0.36	0.98	0.27	I	34	P4I4	0.36	0.82	0.64	I	6	P1I6	-0.45	0.85	1.40	I	0.45
38	P4I8	-0.55	1.28	1.17	I	17	P2I7	0.36	1.19	0.74	I	23	P3I3	-1.32	1.14	1.56	I	0.60
39	P4I9	-0.40	0.99	1.34	I	48	P5I8	0.36	1.28	1.77	I	10	P110	-0.66	1.07	1.62	I	0.57
40	P410	0.11	1.20	0.89	I	56	P6I6	0.61	2.02	2.54	I	31	P4I1	-1.25	1.16	1.67	I	0.61

BEST COPY AVAILABLE

Table 7.3 (Continued)

SERIAL ORDER						DIFFICULTY ORDER						FIT ORDER						
SEQ NUM	ITEM NAME	ITEM DIFF	DISC INDX	FIT MN SQ	I I	SEQ NUM	ITEM NAME	ITEM DIFF	DISC INDX	FIT MN SQ	I I	SEQ NUM	ITEM NAME	ITEM DIFF	DISC INDX	FIT MN SQ	I I	POINT BISER
41	P5I1	[illegible]	1.11	1.02	I	60	P6I10	0.65	1.21	2.03	I	28	P3I8	-0.94	1.31	1.77	I	0.68
42	P5I2	[illegible]	1.22	0.99	I	58	P6I8	0.69	1.65	1.87	I	48	P5I8	0.36	1.28	1.77	I	0.65
43	P5I3	1.11	0.84	0.65	I	49	P5I9	0.77	1.03	2.56	I	27	P3I7	-0.31	1.32	1.83	I	0.68
44	P5I4	0.88	0.45	6.10	I	33	P4I3	0.77	1.27	0.41	I	5	P1I5	-1.19	1.06	1.84	I	0.55
45	P5I5	1.23	0.74	3.26	I	32	P4I2	0.84	1.14	0.82	I	58	P6I8	0.69	1.65	1.87	I	0.72
46	P5I6	1.92	0.96	3.11	I	44	P5I4	0.88	0.45	6.10	I	13	P2I3	-0.25	0.92	1.91	I	0.52
47	P5I7	2.00	0.89	3.94	I	54	P6I4	1.04	1.16	2.19	I	60	P6I10	0.65	1.21	2.03	I	0.60
48	P5I8	0.36	1.28	1.77	I	43	P5I3	1.11	0.84	0.65	I	22	P3I2	-1.32	0.90	2.17	I	0.46
49	P5I9	0.77	1.03	2.56	I	19	P2I9	1.15	0.59	5.23	I	8	P1I8	-1.56	0.99	2.18	I	0.53
50	P5I10	1.73	0.92	1.22	I	45	P5I5	1.23	0.74	3.26	I	54	P6I4	1.04	1.16	2.19	I	0.59
51	P6I1	1.77	-45.78	9.07	I	55	P6I5	1.27	1.24	1.12	I	56	P6I6	0.61	2.02	2.54	I	0.73
52	P6I2	2.00	0.74	1.03	I	53	P6I3	1.42	1.38	1.08	I	49	P5I9	0.77	1.03	2.56	I	0.59
53	P6I3	1.42	1.38	1.08	I	35	P4I5	1.46	0.38	6.98	I	46	P5I6	1.92	0.96	3.11	I	0.43
54	P6I4	1.04	1.16	2.19	I	50	P5I10	1.73	0.92	1.22	I	45	P5I5	1.23	0.74	3.26	I	0.40
55	P6I5	1.27	1.24	1.12	I	51	P6I1	1.77	-45.78	9.07	I	47	P5I7	2.00	0.89	3.94	I	0.38
56	P6I6	0.61	2.02	2.54	I	46	P5I6	1.92	0.96	3.11	I	19	P2I9	1.15	0.59	5.23	I	0.43
57	P6I7	2.29	0.92	1.20	I	47	P5I7	2.00	0.89	3.94	I	44	P5I4	0.88	0.45	6.10	I	0.33
58	P6I8	0.69	1.65	1.87	I	52	P6I2	2.00	0.74	1.03	I	35	P4I5	1.46	0.38	6.98	I	0.26
59	P6I9	2.42	0.95	0.40	I	57	P6I7	2.29	0.92	1.20	I	1	P1I1	-1.19	-13.94	8.98	I	0.17
60	P6I10	0.65	1.21	2.03	I	59	P6I9	2.42	0.95	0.40	I	51	P6I1	1.77	-45.78	9.07	I	0.37
	MEAN	-0.00	0.02	1.73														
	S.D.	1.20	6.32	1.85														

CORRELATION DIFF*DISC= -0.15 DIFF*MNSQ= 0.27 DISC*MNSQ= -0.65

of the square root of two over its degrees of freedom which will vary between 1 and 5 depending on the number of score groups defined by the analysis. The three correlations of difficulty with discrimination, difficulty with fit mean square and discrimination, difficulty with fit mean square and discrimination with fit mean square have zero expectations" (Wright and Mead, 1975, p. 19).

Figures 7.3, 7.4, 7.5, and 7.6 display the following plots, respectively: (1) item z^2 against the probability of a person in an ability group answering the item correctly, (2) item fit mean square against item difficulty, (3) item fit mean square against the index of item discrimination, and (4) item discrimination index against item difficulty. These figures complete the detailed analysis of the fit of the data on MCC Form 14 to the Rasch model.

By reference to the third column in Table 7.3, it will be observed that there are several items that do not fit the Rasch model. Two of these items are numbered 35 and 47. Their high "Fit Mean Square" indices suggest they are not operating as expected or like the other items. Item 35 required the students to reconstruct the following sentence:

Everyone was bargaining _______________ back and forth.

The possible alternative choices for the deleted word included:

a. snugly
b. merrily
c. loudly
d. honestly
e. painfully

The correct answer was "loudly." The alternatives, according to design specifications, were to have been semantically implausible. In item 35, however, "merrily" and "honestly" are both semantically plausible and there is nothing in the passage to suggest that "loudly" was more appropriate than "merrily"

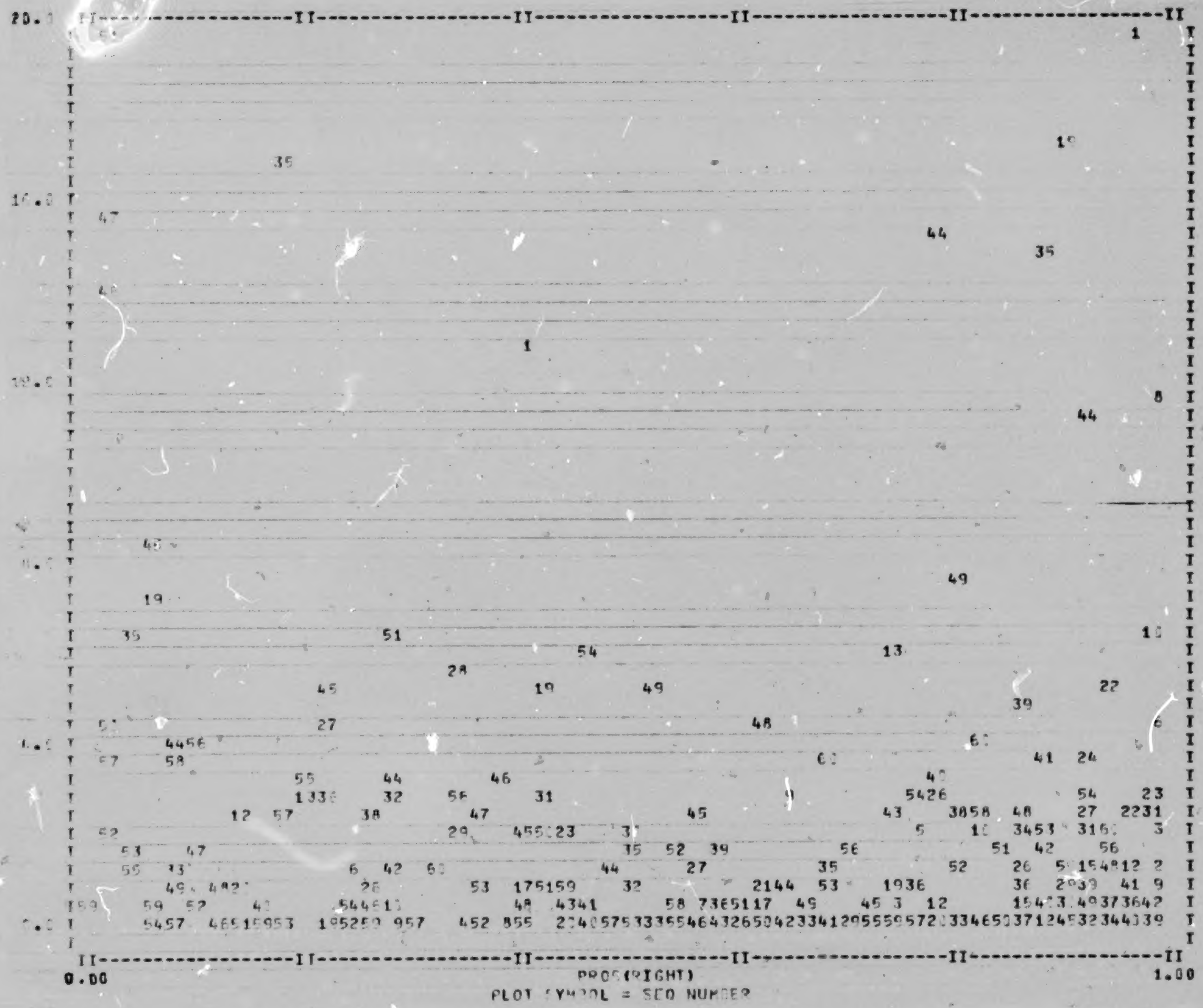


Figure 7.3. NYSSED Multiple-Choice Cloze Study July 1975 Analysis of Form 14 Item Z Square (Y) Versus Prob (RIGHT) (X).

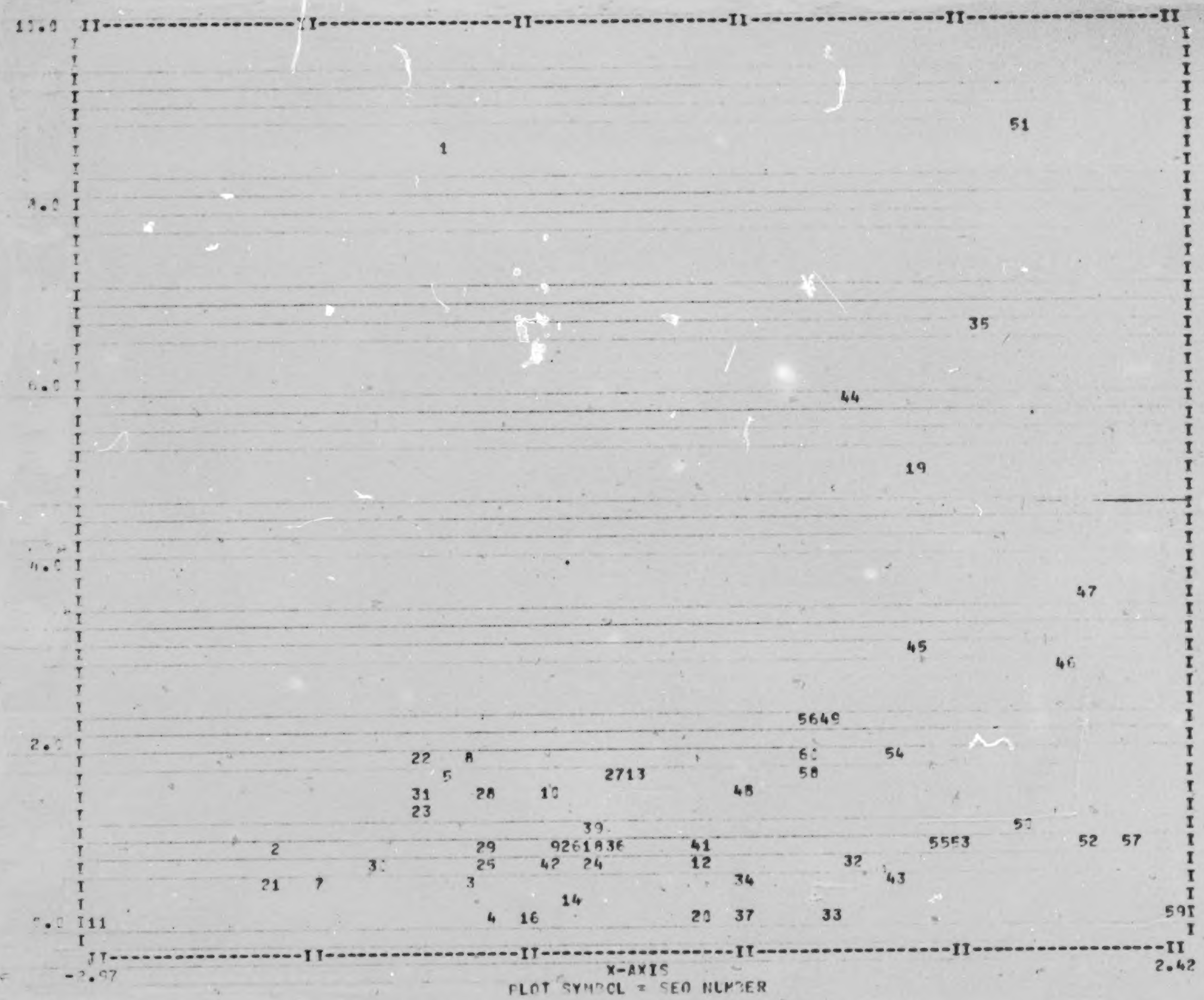


Figure 7.4. NYSSED Multiple-Choice Cloze Study July 1975 Analysis of Form 14 Fit Mean Square (Y) Versus Difficulty (X).

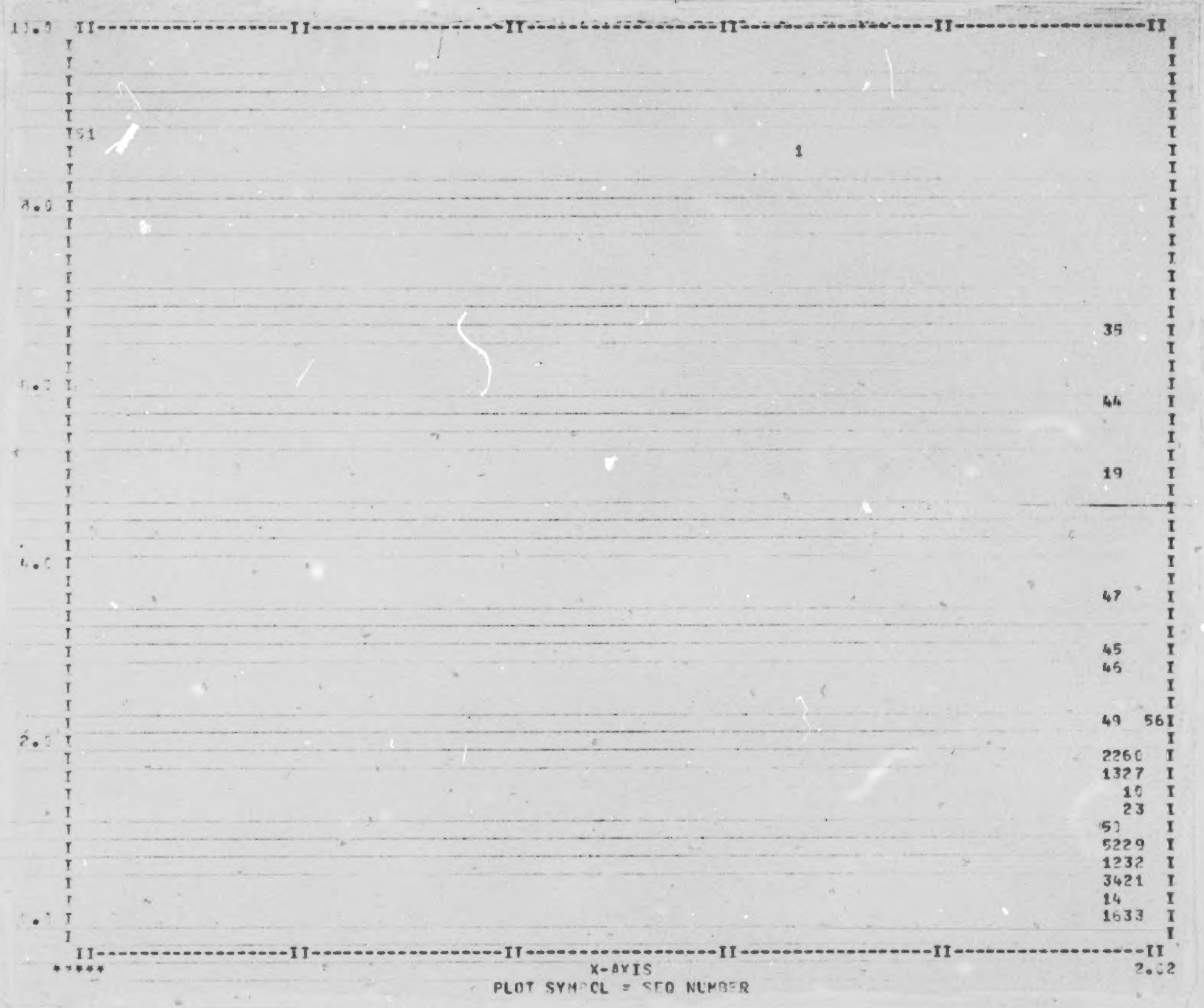


Figure 7.5. NYSSED Multiple-Choice Cloze Study July 1975 Analysis of Form 14 Fit Mean Square (Y) Versus Discrimination (X).

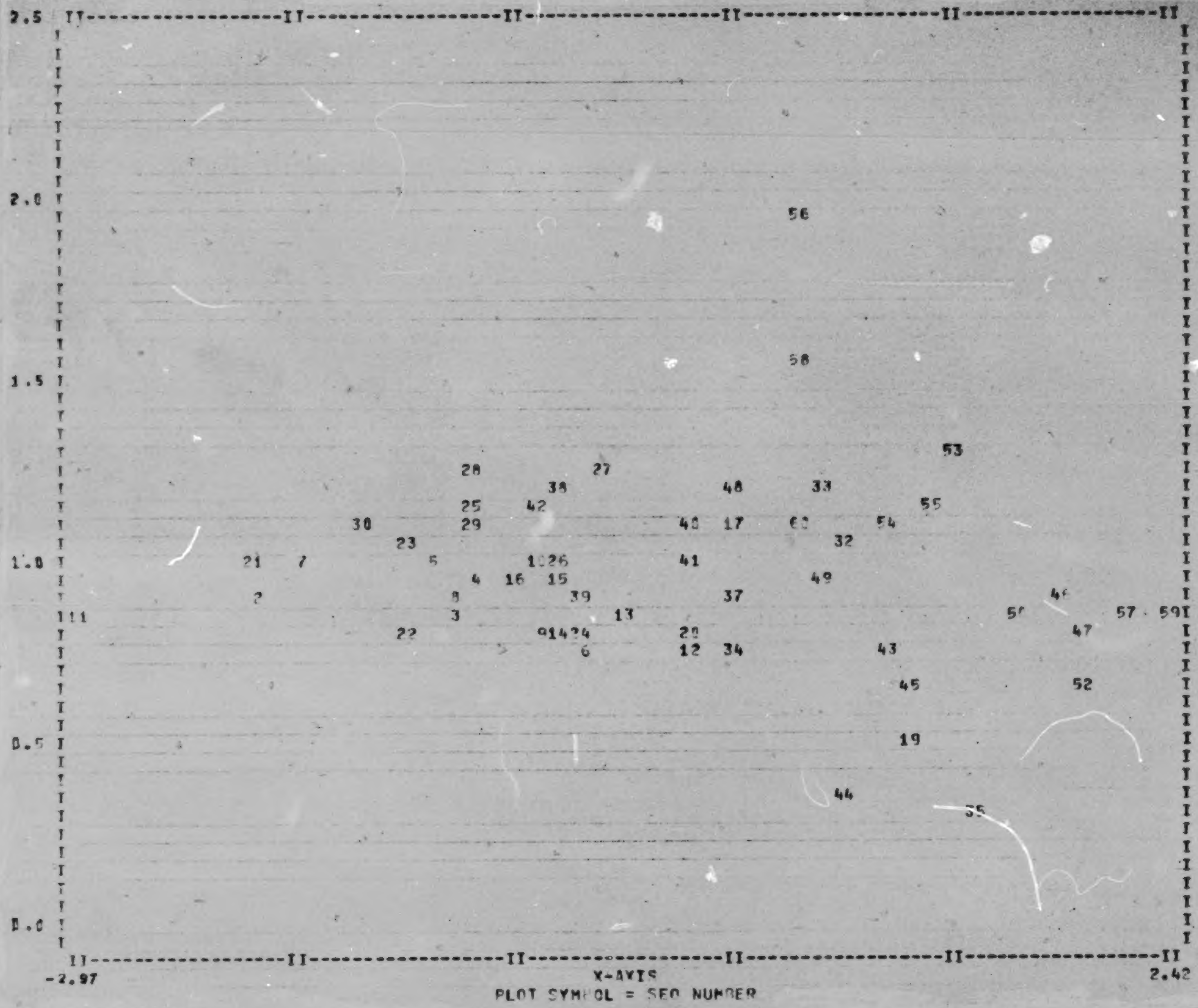


Figure 7.6. NYSSED Multiple-Choice Cloze Study July 1975 Analysis of Form 14 Discrimination (Y) Versus Difficulty (X).

or even "honestly," though it seems incongruous with "bargaining." This fault in the item undoubtedly forced it to operate at odds with the rest of the test.

Item 47 required the students to reconstruct the following sentence:

Neither ___(race)___ pleased the gods ______________.

The alternative choices included:

a. obediently
b. entirely
c. miraculously
d. modestly
e. incompetently

The explanation of the misfit of this item does not reside in the distractors, which fulfill the condition of semantic implausibility. Rather, the sentence upon which this item was based seems to have presented two problems. First, it contained another item which was very difficult, probably because the word deleted--race--is used in a somewhat unfamiliar way (i.e., to refer to elves and dwarfs, the subjects of the passage). Undoubtedly, the difficulty of this item contributed to the difficulty of the following, or misfitting, item. Further, it seems likely that a more typical rendition of the sentence containing the misfitting item would have placed the adverb (entirely) between the subject and the verb, rather than terminally. Possibly the difficulty of the previous item and the somewhat uncommon sentence structure combined to produce difficulty and confusion.

The calibration of the items in Form 14 would proceed by eliminating those students who seem to be causing most of the misfit with the Rasch model. In most cases, students who perform at the extremes of ability distributions cause the greatest misfit. After removing the student with extreme ability characteristics, the fit of items to the model usually improves. When all of the items fit the model, the general structure of the test would no longer be suspect. The resulting item difficulties or average passage difficulties would thus be placed on an equal-interval scale from low to high difficulty. This technique for scaling the passages without reference to the students taking the test, provides the basis for placing all of the passages, across content domains, on the same difficulty scale. A complete calibration design network must be developed to guarantee that all items, and thus all passages, are calibrated on a common scale.

Rasch Calibration on a Common Latent Variable

As previously noted, our major interest in calibrating items is the placement of items on a common scale for a latent variable called literal comprehension. Speaking in operational terms, "When the pool of items from which we select the elements for a best possible test has been calibrated on a latent variable, then these items and their locations on the latent variable provide its operational definition. A measurement of a person on the variable will place him among items with difficulties near his estimated ability. The meaning of his position on the variable will be defined by these nearby items" (Wright and Douglas, 1975, p. 4). Therefore, if one is interested in measuring a person's ability to literally comprehend written discourse, one could develop and calibrate items on the hypothesized trait, then use the items to estimate a person's ability on that latent trait. From a measurement perspective, the best estimate of that ability would come from items calibrated on a common, equal-interval, zero-point scale. This type of scale can be developed with existing methodologies.[6] Calibration of items or passages on a common scale would include the following major tasks:

---

[6] Personal communication with Benjamin D. Wright and Ronald Mead, 1975.

1. Define conceptually the variable or trait under investigation;

2. Choose a test or passage/item format requiring skills specified in the conceptual definition;

3. Prepare a passage/item pool;

4. Develop a calibration plan;

5. Construct calibration tests;

6. Administer the calibration tests;

7. Collect calibration data and analyze with Wright-Mead CALFIT computer program;

8. Select items from item-analysis results based on discrimination and fit-mean-square statistics;

9. Calibrate selected items based on difficulty estimates;

10. Calibrate passages based on average passage difficulty;

11. Determine the effect a particular content domain might have on the difficulty of items;

12. Assuming item and passage difficulties are "content free," chain or link all passage calibrations onto one underlying scale with equal intervals and a meaningful zero point.

This calibration plan will result in the placement of all cloze passages in the Test Development Notebook (TDN) on a common scale. Each passage will be assigned a difficulty index that can guide teachers in the selection of those passages that would be most appropriate for testing their students. Tasks 1 through 4 have been completed. Over 1,000 passages, in multiple-choice cloze format, are available for administration and calibration. These passages cover reading content in the following domains:

1. Textual Materials in Reading, Language Arts, Social Studies, Science, and Mathematics;

2. Citizen Material from newspapers and magazines;

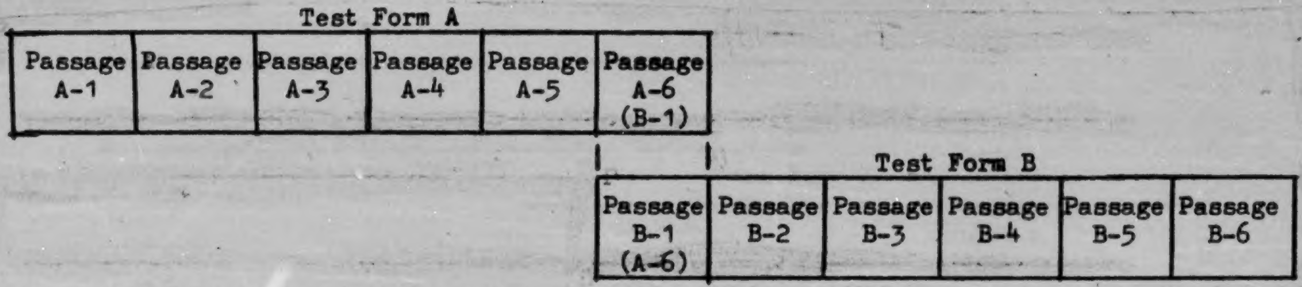
3. Consumer Materials from catalogs, advertising, instructions, and so forth; and

4. Reference Material from test instructions, children's magazines, encyclopedias, and so forth.

In addition, the calibration plan will include the 300 multiple-choice passages that have been developed for the Test Development Notebook (TDN).

The calibration plan. Item/passage calibration on a common variable involves careful planning, complex data management, and extensive test analysis. The design to be presented will require the construction and administration of more than 400 separate test forms. These forms will be administered to over 50,000 students. In order to place each passage on a common scale, a data management system must be developed to monitor the estimation and re-estimation of the difficulties of 2,500 passages (including some redundancies) as they appear on different test forms.

With the Rasch basic measurement model, calibrating items or passages on a common scale is possible when the same items or passages are placed on two separate test forms. It is best when one of the test forms is slightly more difficult than the other. This approach links two test forms together through common items or passages. For example, define two test forms, A and B, each containing 6 multiple-choice cloze passages with 10 items each. On Test A the passages increase in difficulty to the 6th passage. On Test B, passage 1 is identical to the 6th passage in Test A, is followed by 5 more passages of increasing difficulty. The basic linking model can be expressed graphically as follows:


Test Form A
Passage A-1
Passage A-2
Passage A-3
Passage A-4
Passage A-5
Passage A-6 (B-1)
Test Form B
Passage B-1 (A-6)
Passage B-2
Passage B-3
Passage B-4
Passage B-5
Passage B-6

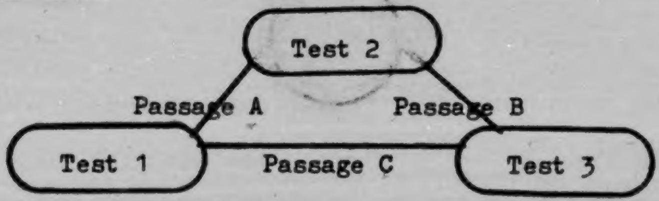
In a calibration plan, one sample of students would respond to Test Form A, and a similar sample of students would respond to Test Form B. The two tests would then be calibrated separately using the Rasch model in the Wright-Mead CALFIT computer program. The resulting calibrations would provide two scales with equal measurement units but different origins. The passage difficulties in Test Form B can be placed on the same scale (with the same origin) as Test Form A by adding a translation constant to all of the item difficulties on Test Form B. The translation constant is calculated as follows:

> Define the average Rasch difficulty for the items in passage A-6 as d_{A-6} and d_{B-1} for passage B-1, then the translation constant from Test Form B to Test A is defined as follows:
>
> $$t_{BA} = d_{A-6} - d_{B-1}$$

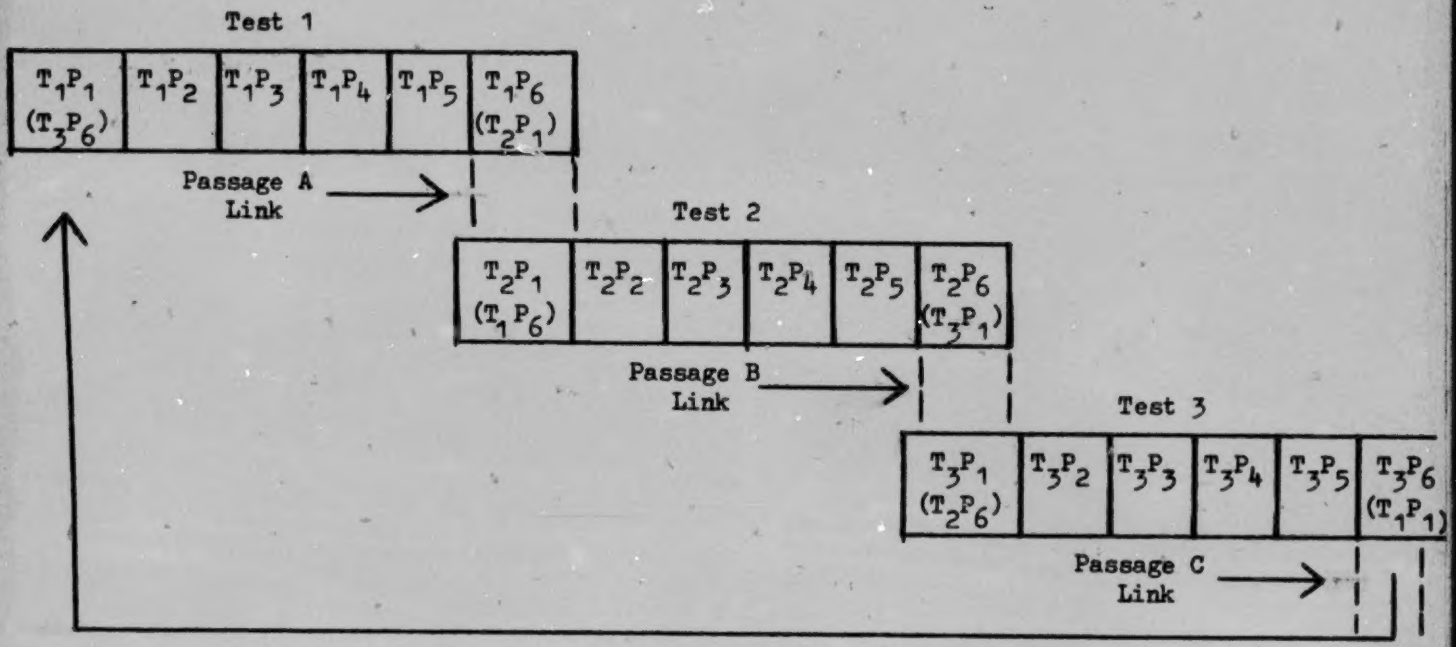
The average passage difficulties in Test B can thus be placed on the scale of Test Form A by adding the translation constant, t_{BA}, to all of the average passage difficulties calculated for Test Form B. That is, $d_{B-1(A)} = d_{B-1} + t_{BA}$, $d_{B-2(A)} = d_{B-2} + t_{BA}$, etc.

This form of elementary linking will place all passage difficulties on Test Form B on a common scale with Test Form A. However, there is no possibility of cross-checking the stability of these translations to a common scale, at least not within this elementary design.

A slightly more complex linking scheme allows for cross-checking the placement of passages on a common scale. The more complex scheme may be diagramed as follows:



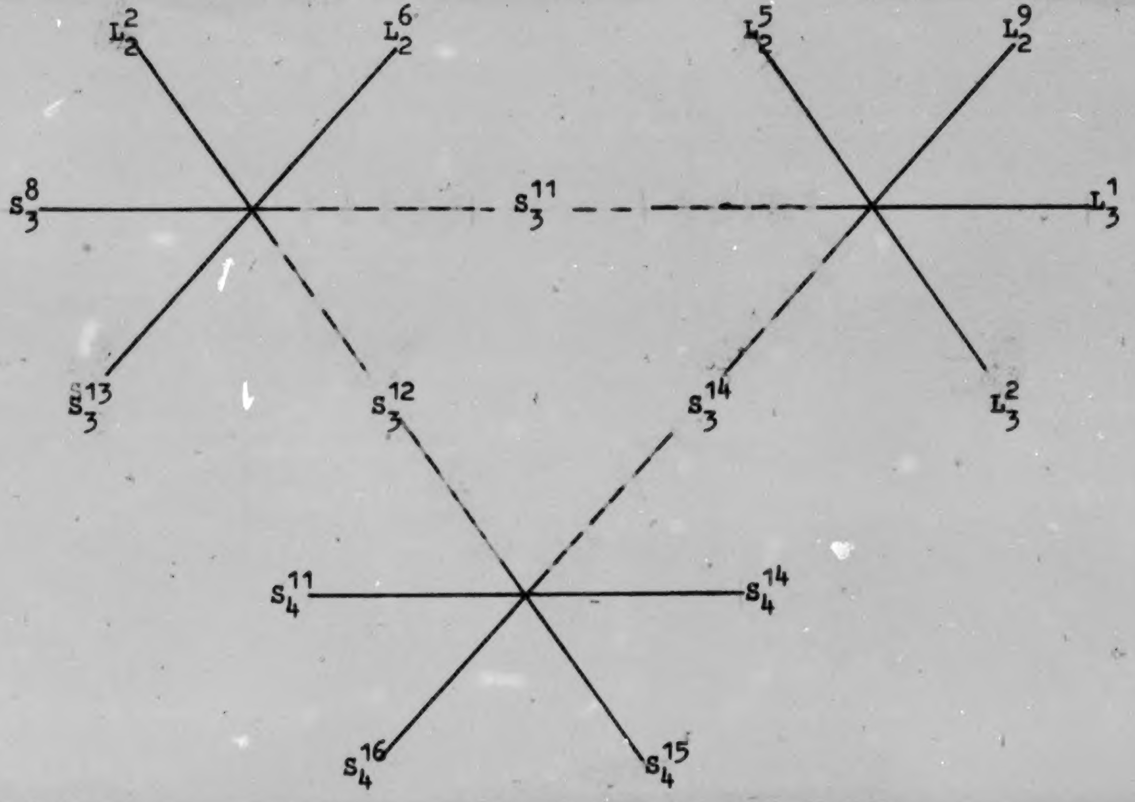
In this design, passages A, B, and C are common to two tests. (This basic, triangular pattern will form the core of the final calibration design.) These common relationships can also be noted as:


Test 1
T1P1 (T3P6)
T1P2
T1P3
T1P4
T1P5
T1P6 (T2P1)
Passage A Link
Test 2
T2P1 (T1P6)
T2P2
T2P3
T2P4
T2P5
T2P6 (T3P1)
Passage B Link
Test 3
T3P1 (T2P6)
T3P2
T3P3
T3P4
T3P5
T3P6 (T1P1)
Passage C Link

The translation constants from test to test are denoted as t_{A21}, t_{B32}, and t_{C13}. These translation constants are calculated the same way they were for the basic link, that is, $t_{A21} = d_{T_1P_6} - d_{T_2P_1}$, $t_{B32} = d_{T_2P_6} - d_{T_3P_1}$, and $t_{C13} = d_{T_3P_6} - d_{T_1P_1}$.

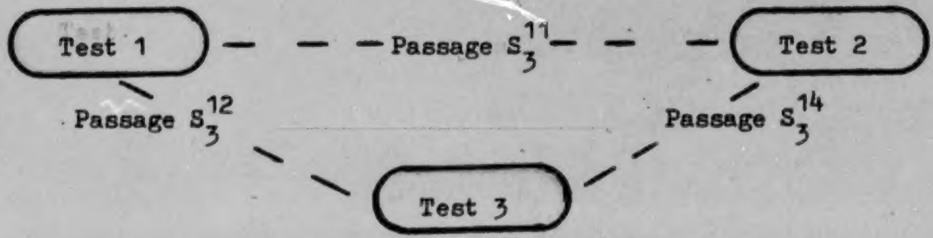
With this basic, triangular, cross-checking calibration design, an estimate of the consistency with which passages are placed on a common scale is obtained from the expected value of the sum of the translation constants. That is, $t_{A21} + t_{B32} + t_{C13} = 0$. The expected value of the sum of the translation constants is zero. When there is a deviation from zero, an adjustment is made to equalize the relative values of the translation constants, while maintaining their sum at zero. This triangular unit for estimating the translation constants is the basic unit of analysis to be used in the complete calibration design or network.

The application of the triangular unit of analysis in the final calibration design is illustrated in Figure 7.7. This is a small portion of the entire final calibration design. Each six-pointed star is a complete test form. (Each test form will be administered to approximately 175 students.) For illustrative purposes, an enlargement of a portion of the network follows:



L_2^2
L_2^6
L_2^5
L_2^9
S_3^8
S_3^{11}
L_3^1
S_3^{13}
S_3^{12}
S_3^{14}
L_3^2
S_4^{11}
S_4^{14}
S_4^{16}
S_4^{15}

Here the triangular unit of analysis, previously discussed, is marked by dashes. The unit of analysis for linking these passages together would be:



Test 1
Passage S_3^{11}
Test 2
Passage S_3^{12}
Passage S_3^{14}
Test 3

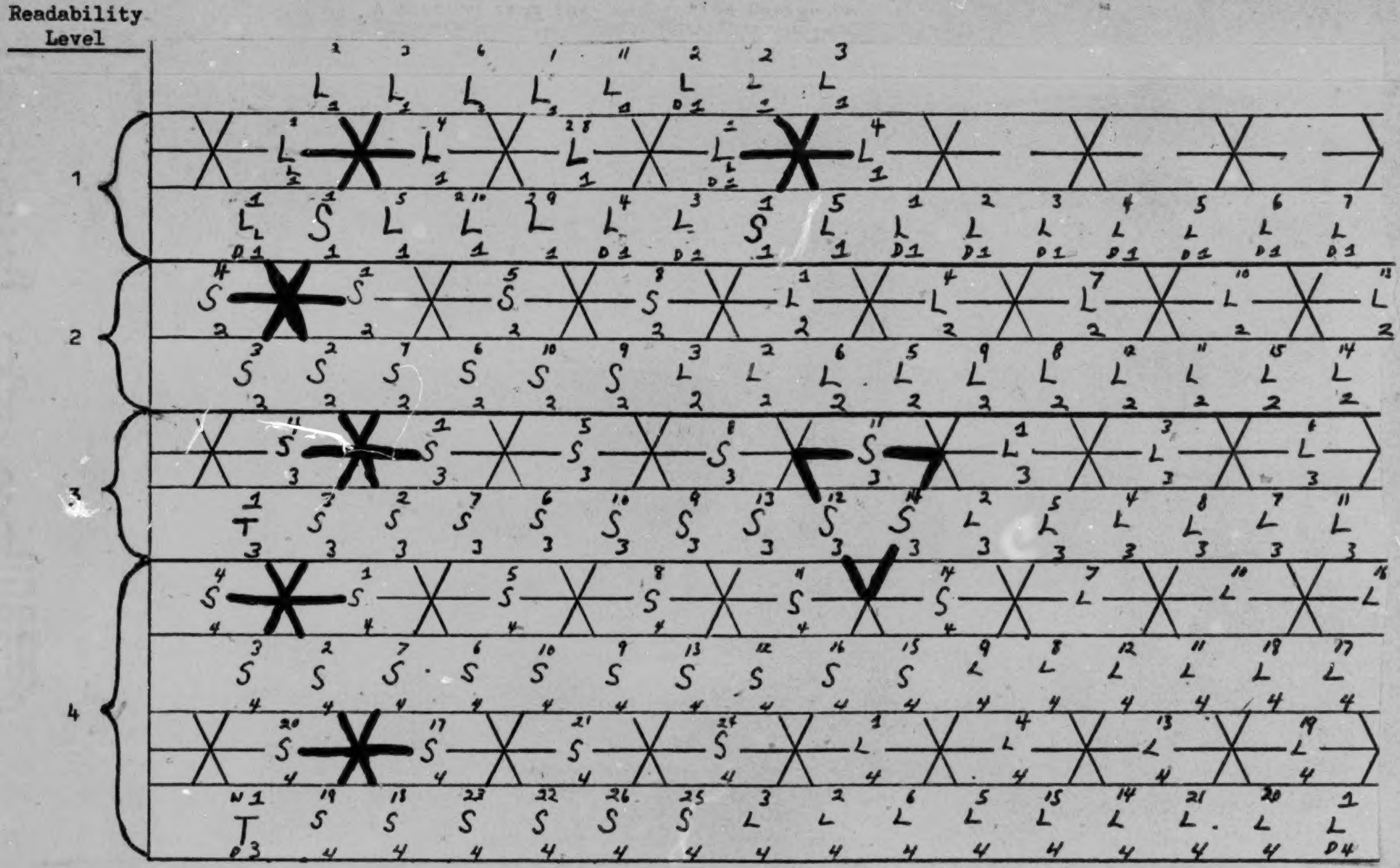

Readability Level
1
2
3
4

Figure 7.7 A section from the calibration design network.

Passages S_3^{11}, S_3^{14}, and S_3^{12} are common to two test forms. The translation constants from test to test would be calculated as follows:

$$t_{S_3^{11}21} = d_{T_1P_6} - d_{T_2P_1}$$

$$t_{S_3^{14}32} = d_{T_2P_6} - d_{T_3P_1}$$

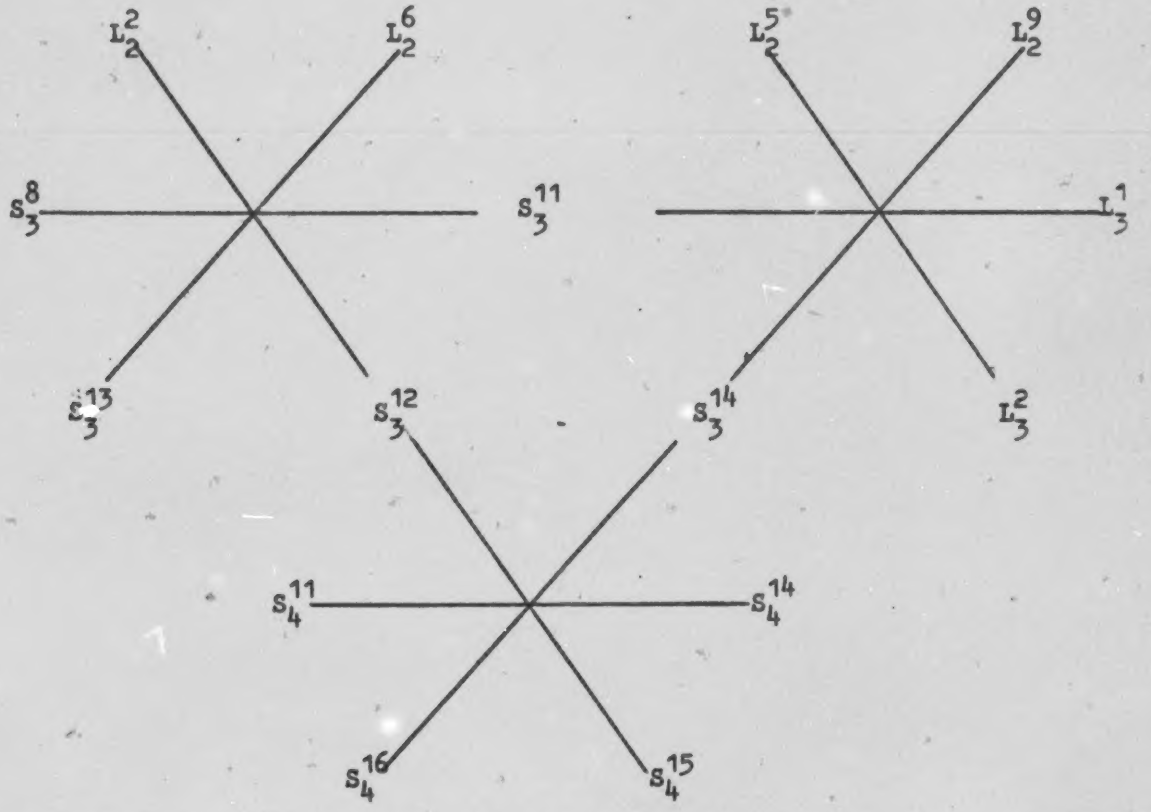
$$t_{S_3^{12}13} = d_{T_3P_6} - d_{T_1P_1}$$

These translation constants must also sum to zero. If they do not, minor adjustments will be made in the individual translation constants in order to equalize them. (Thousands of calculations must be made to translate every passage in the network onto a common scale.)

The complete calibration design will involve the calibration of all passages in the Test Development Notebook onto a common scale. The portion of the calibration network illustrated in Figure 7.7 refers to actual passages from particular content domains. The "network passage codes" for the particular content domains and sub-domains are as follows:

Textual Domains	Network Passage Code
1. Reading/Literature	L
2. Language Arts	T
3. Social Studies	S
4. Science	I
5. Mathematics	A
Citizen Domain	
6. Newspapers	N
7. Magazines	M
Consumer Domain	C
Reference Domain	R

The sample portion of the original calibration design network in Figure 7.7 illustrates the use of the passage codes. It includes three tests that will span three readability levels, namely, Levels 2, 3, and 4. For convenience the example is repeated here.



The first test in the upper left-hand corner of the example is composed of 2 Reading/Literature (L) passages--L 2,2 and L 6,2, and 4 Social Studies (S) passages--S 8,3; S 13,3; S 12,3; and S 11,3. The exact classification of each passage by identification number, readability level, and duplication (i.e., whether the passage is being used more than once) is denoted as follows:

$$\text{(Duplicate) D}\ \boxed{L}\ \begin{matrix} \text{2(I.D. Number)} \\ \text{2 (Readability Level)} \end{matrix}$$

This would indicate passage number 2 at readability level 2 and the fact that it is being used as a duplicate to complete this test form.

As mentioned previously, a data management system will be used to monitor the manipulations and calculations necessary to place all passages on a common scale. This system will include the following major operations:

1. Calculate a complete Rasch analysis on each form;
2. Average the Rasch difficulties for each passage within each test form;
3. Merge identification numbers with each passage from the complete calibration design, or network;
4. Calculate all possible translation constants indicated by the identification numbers;
5. Equalize the translation constants within each triangular unit of analysis such that the impact on adjacent translation constants is minimized, at least within a difficulty level or unique content area;
6. Determine the passage with the lowest possible difficulty, and begin to translate passages from separate forms onto this original difficulty scale; continue the translation throughout the calibration network until all passages are calibrated on the same equal-interval scale with a meaningful zero point.

This calibration design assumes that changing from one content area to another will not affect the measurement of literal comprehension with the multiple-choice cloze test. Due to the complex interrelationships among passages in the present calibration design, it will be possible to test the effect of content area (e.g., science versus math) on the measurement of literal comprehension. There is little possibility that there will be a "content effect." If there is, the passages will be recalibrated within their respective content domains and organized accordingly within the Test Development Notebook.

Implementation plan. The implementation plan for the calibration design will parallel the validation plan but will require considerable "front-end" work in the areas of planning, computer programming, computer-managed data base and monitoring systems. The complexity of the final calibration design will be more manageable, pending considerable research and development via an experimental implementation of a reduced calibration design.

The time line for implementing the calibration design will be coordinated with the time line for the validation and productivity plan. For convenience, the overall research and development time line is provided again in Figure 7.8. The three major research components (i.e., validation, calibration and productivity) will be implemented sequentially, with overlapping phases.

A more detailed time line for the calibration research component is presented in Figure 7.9. This figure summarizes the implementation and completion of major tasks, the delivery of products, and the assistance of the external review panel. The major tasks include:

1. Planning with input from appropriate consultants of the external review panel;

2. The immediate development and implementation of a pilot study of the feasibility of the experimental calibration design network;

3. A major effort to modify the CALFIT computer program to conform its design with data management needs;

4. The design and up-keep of a data management and filing system that is mostly computerized;

5. Finalization of the complete calibration design network, including improvements based on information synthesized from the pilot study;

6. The preparation of 450 or more unique test forms based on the calibration design network;

7. Implementation of a major data-collection effort based on the calibration design network with 150 to 200 students responding to each test form;

8. Major data analysis and calibration of individual passages including a determination of a possible "content effect" and the placement of each passage on a common, equal-interval, "zero-point" scale;

9. The calculation of derived scores that will improve the interpretation of test scores at the local level;

10. Completion of a final report on the calibration of all multiple-choice cloze passages in the Test Development Notebook, including a delineation of the use of derived

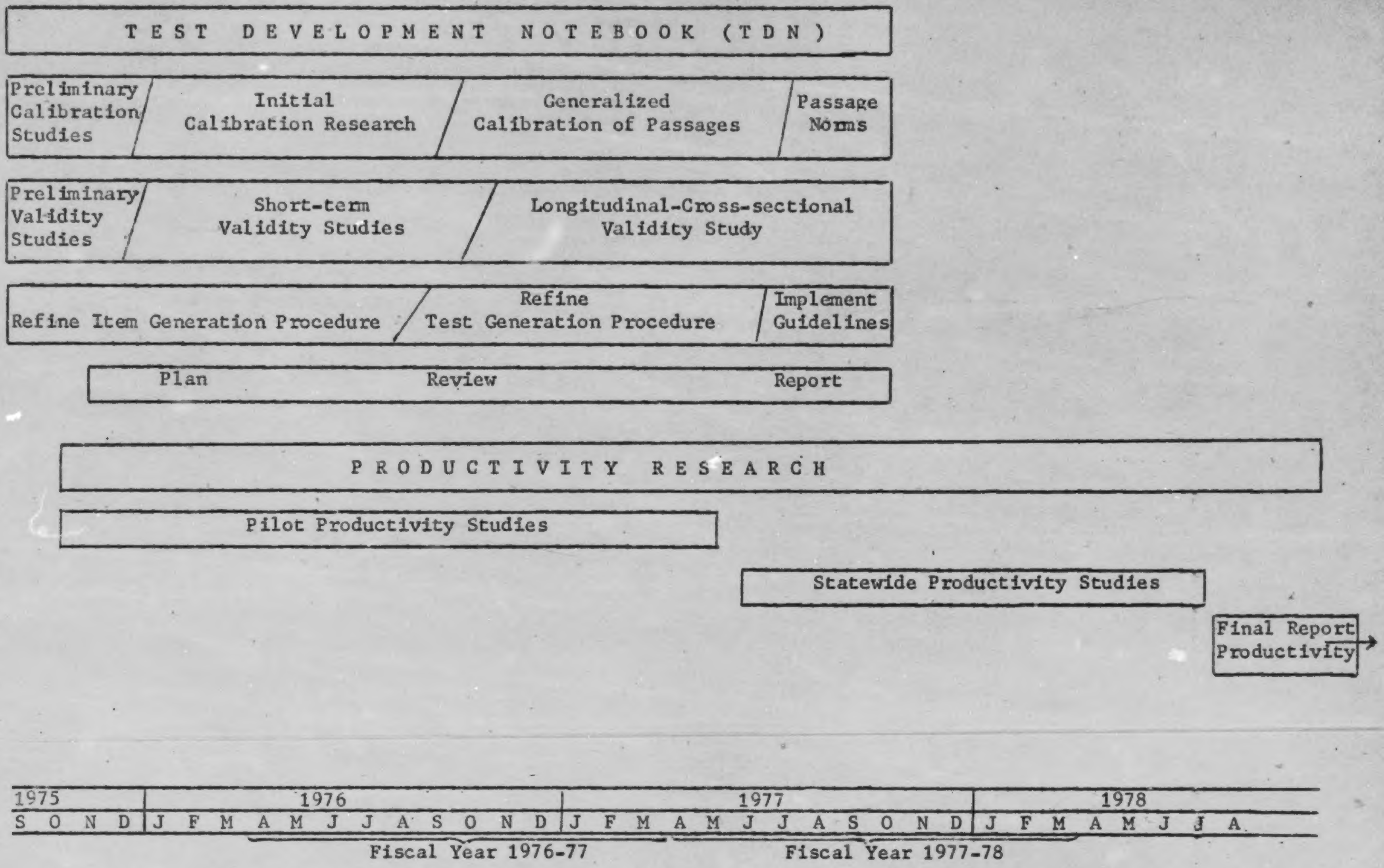


Figure 7.8. Research and development activities.

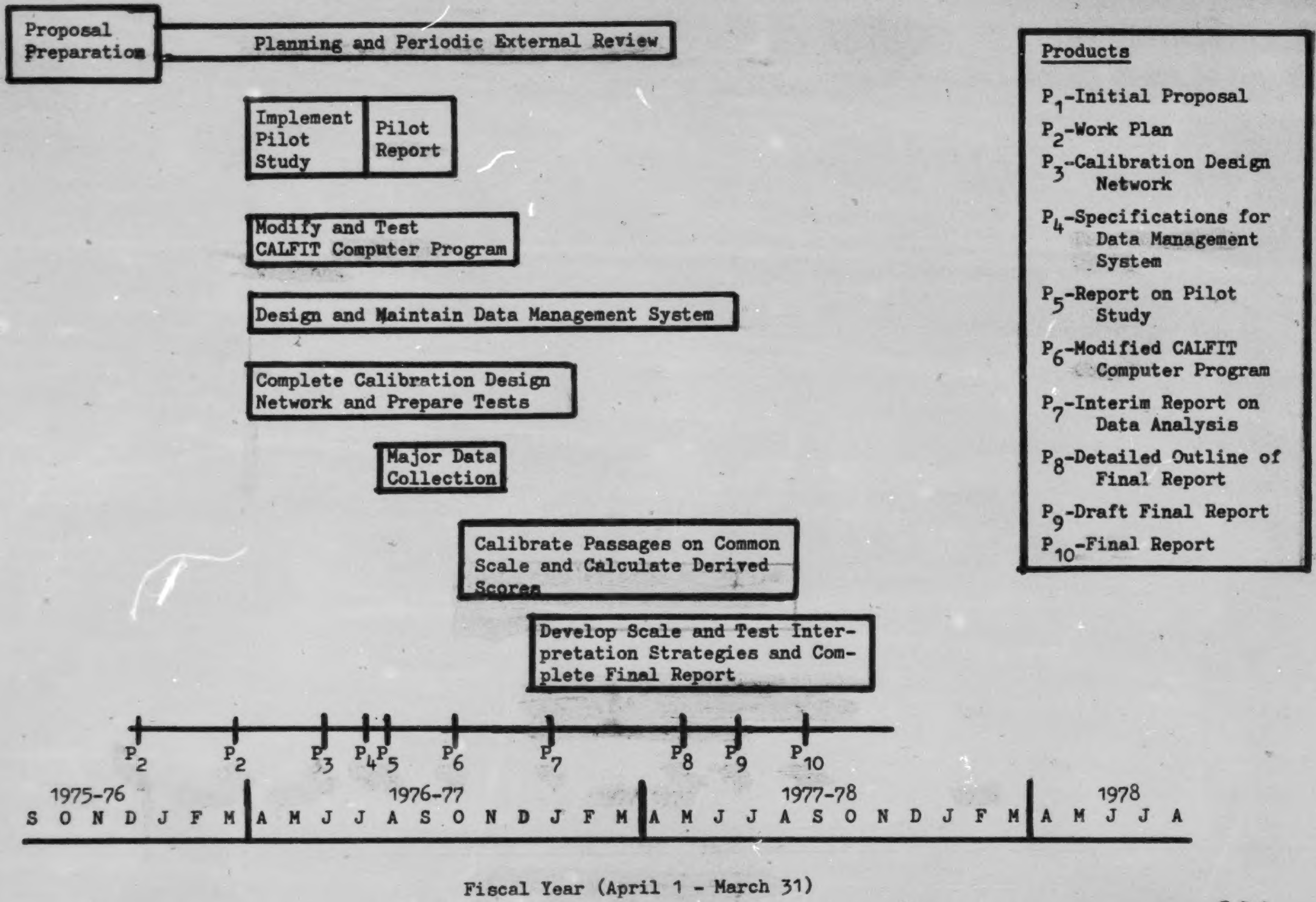


Figure 7.9. Timeline by tasks for implementing calibration plan.

scores to improve the interpretability of test scores at the local level.

The schedule for completion of major products during the implementation of the calibration plan is set forth in Figure 7.9.

As noted, the calibration plan will be directed from the Bureau of School and Cultural Research and coordinated with the validation plan for the multiple-choice cloze test.

CHAPTER VIII

ANALYSIS OF CLOZE PASSAGES AND ITEMS

This chapter describes three related but distinct phases of an analysis of the multiple-choice cloze exercises administered during May and June 1975 to approximately 5,000 students in an urban school district in upstate New York. As indicated in Chapter V, the cloze exercises consisted of three levels of test forms: Level I, administered to students in grades 1, 2, and 3; Level II, administered to students in grades 4, 5, and 6; and Level III, administered to students in grades 7, 8, and 9. At each level there were 12 separate test forms, for a total of 36 test forms. All passages on the 36 test forms were taken from the reading/literature domain. Each test form contained six passages and, with exceptions at Level I, each passage was accompanied by 10 multiple-choice items. Generally, then, each test form featured 60 multiple-choice items. Since the majority of students taking the Level I tests were first and second graders, the Level I test forms each contained three shorter passages with fewer items (i.e., either three or five items). Thus, the Level I test forms vary from the general pattern in that they contain either 39 or 41 items, instead of 60.

Passages were randomly selected in order to assure that all test forms at a given test level (i.e., Levels I, II, or III) would be as nearly equivalent as possible. The passages on every test form were arranged in ascending order of readability (i.e., from the least to the most difficult). Each test form at a given level, then, represented the same range of reading difficulty.

Utility of Analysis

The 36 cloze test forms and the data derived from their administration are the materials upon which the analysis reported in this chapter is based. Since the readability levels of all cloze passages were determined by the Spache and Dale-Chall readability formulas, and since the formats and items were governed by the same constructional procedures, any sampling of cloze materials should be representative of the entire corpus.

The analysis here reported represents the first systematic inspection of the cloze materials. It affords the first opportunity of determining how successfully the cloze materials conform to expectations. The analysis had two basic purposes: One was to determine the need for modification of the cloze format and procedures to assure maximum consistency and objective reproducibility of passages and items; the other was to study the effectiveness of the Spache and Dale-Chall readability formulas in ranking passages by difficulty in order to provide preliminary guidelines for the selection of passages for test assembly.

Phases of Analysis

The first phase of the analysis was a critical examination of the cloze test forms largely completed prior to availability of test result data but informed where necessary by the data as it became available. It was intended to discover and categorize ostensible flaws in passages and items which might either represent inconsistent application of cloze procedures or predict deviancies in the behavior of passages and items on tests. The second phase of the analysis examined the effectiveness of readability formulas in ranking passages by difficulty. This involved the calculation (using Rasch analysis data) of passage easiness scores by grade and across grades for each passage at each test level, and the inspection of these scores both by test form and by designated readability level. The third phase of the analysis featured an inspection of items on 12 test forms (four at each testing level) aimed at identifying and explaining deviant items in order to determine the need for further revisions of cloze procedures, to suggest modifications which might improve cloze materials, and to make preliminary observations on the unidimensionality of the cloze materials.[1] Though these three analytical phases are substantially interdependent and mutually informative, they will be described separately for the sake of simplification and clarity.

Phase 1--Critical Examination of Cloze Passages and Items

This phase of the analysis was initiated as a review of the 36 cloze test forms for the sake of assuring consistency among the items. Two steps were involved. The first step was the actual review of every item on every test form. This step resulted in the identification of a number of items which were flawed either in terms of violations of existing cloze rules or procedures or in terms of desirable revisions in the cloze procedures which came to light during the review process. The second step in this phase of the analysis involved determining the extent to which flawed items were able to predict items subsequently revealed as deviant by the test result data.[2]

---

[1] These observations involved inspection of misfitting items. Items with high fit mean squares on the Rasch measurement model are identified as misfitting or as measuring characteristics other than those which the items were intended to measure. Misfitting items on the cloze tests are assumed to be measuring some trait other than or in addition to literal comprehension.

[2] The criteria for identification of deviant items are presented in Phase 3.

Step 1--Review Procedures and Results

Distractors for every item on every test form were inspected in context by at least two and occasionally three reviewers. Each reviewer would independently examine every passage on a test form, for every item testing each distractor in the space left in the passage by the deleted word, and noting items which seemed flawed. Then at least two reviewers would examine the test form together, discussing items with possible flaws and listing in categories all items thus identified as flawed.

The four categories of flawed items identified by this review include grammatically (syntactically) implausible distractors, semantically plausible distractors, idioms, and errors.

Grammatically (syntactically) implausible distractors. Cloze procedures require that distractors be taken from part-of-speech lists identical to the part of speech (i.e., function in context) of the deleted word. Many of the flawed items in this category resulted from the use of an inappropriate part of speech list. Others had distractors disagreeing with deleted verbs in tense or number or distractors disagreeing with determiners (a, an, the) preceding deleted words. But the most frequent problem was unworkable distractors taken from the appropriate part-of-speech list. That is, every word which can function as the same part of speech as a deleted word cannot necessarily function in the same contextual position as that deleted word. For example, if the word town, a noun in the sentence "John went to town last weekend," is deleted, not every possible noun will be grammatically plausible in the position vacated by town. Consider such nouns as house or cow. In this context they are grammatically implausible.

Semantically plausible distractors. Cloze procedures do not permit the use of synonyms as distractors for deleted words. Some synonyms were found among the distractors (e.g., swift, quick). But the majority of these kinds of flawed items involved distractors which, though not synonyms for deleted words, were nonetheless contextually plausible. In the sentence "The children were singing loudly," such distractors as happily or merrily would be plausible, though not synonymous, substitutes for loudly.

Idioms. Defined for cloze procedures as any word for which no grammatically plausible and simultaneously semantically implausible distractors can be found,an idiom is illustrated by the following: "It is any attempt at written communication." Any grammatically plausible substitute for any (e.g., some, one, every) would also be semantically plausible.

Errors. This category of flawed items primarily involved such things as spelling or typographical errors in passages or distractors.

Flawed items amounted to approximately 25% of the total number of items on the 36 test forms. They were distributed among the four categories as follows: grammatically implausible distractors, 58%; semantically plausible distractors, 19%; idioms, 19%; errors 4%.

Three other types of flaws in the cloze materials were noted during the course of the critical review. These involved titles, passage coherence, and violations of cloze rules regarding the number of words between deletions. Titles presented two different kinds of problems. The first is a title which is inappropriate (i.e., either misleading or unrelated) to the passage which it precedes. The second involves a title which cues one or more of the items (i.e., contains one or more of the words deleted from the passage). Passage incoherence is produced by any violent shift in direction or change in topic; sentences not clearly related to or following from one another result in incoherence. The concept of literal comprehension upon which the cloze test is based assumes that apprehension of the meaning carried by a word deleted from a passage depends on a certain percentage of intact surrounding context. Thus, stipulations regarding number of words between deletions were part of the cloze rules or procedures. When the pattern of deletions in cloze passages did not adhere to these stipulations, such violations were noted. Though problems with titles, passage coherence, and deletion patterns were observed and noted, the frequency of their occurrence was not tabulated.

As a result of the first step in the critical review phase of the analysis, two courses of action were initiated: revision of the cloze rules and review of all cloze materials.

The flaws which the first step in the critical review discovered in the cloze test forms revealed some apparent inconsistencies in the application of the cloze procedure. Such inconsistencies, clearly attributable to the experimental and evolutionary nature of the development of the cloze materials, reflected several problems which remained unresolved at the completion of the cloze corpus; to wit, the cloze rules were marred by excessive complexity, numerous exceptions, and insufficient precision.

Review of the test forms having pointed up the necessity for revision of the cloze rules and having identified problem aspects of the cloze materials, revision of the extant cloze rules was begun. After thorough and intensive discussion of the intended nature of the cloze materials and of procedures essential to the achievement or production of such materials, a carefully revised series of cloze rules (i.e., "Rules for Application of the Cloze Procedure," Appendix A) was formulated.

The revised cloze rules are extremely important in several respects. The current rules can be applied with greater assurance of uniformity. In other words, because the current rules are much more efficiently and practically applicable, they assure the production of passages and items with the highest degree of objective reproducibility. Further, the current rules represent an intermediate step crucial to the ultimate development of an algorithm which will permit the maximum degree of computerization of the cloze procedure and, thus, the maximum speed, efficiency, and practicality in the production of cloze materials. Finally, the revised cloze rules have facilitated the implementation of a review of the entire corpus of cloze materials, a review which will greatly enhance the consistency and, thus, the practical utility of the cloze materials.

The review of the cloze corpus now underway is a systematic and thorough one which endeavors to consider every aspect of the cloze passages and items relevant to quality and consistency. Since this review is intended to assure uniform application of the cloze rules, it should effectuate the greatest possible degree of standardization within the cloze passage format and the items. Moreover, since the review is to correct all flawed items and replace those few passages found to be unacceptable in the light of the revised cloze rules, the corpus of cloze materials remaining at the conclusion of the review promises to achieve as nearly as possible that unidimensionality in a testing device which is such a critical need in the measurement of literal comprehension.

Step 2--Flawed Items as Predictors of Statistical Deviance

The first step in the critical examination phase of the analysis of the 36 cloze test forms identified a number of flaws. The following question arose in response to the identification of these flaws: To what extent do the identified flaws anticipate, predict, or explain statistically deviant items? The reviewers' assumptions were that flaws would produce predictable statistical deviance.

Specific predictions. It was expected that grammatically implausible distractors and idioms would produce unusually easy items, because the distractors would be so obviously uncompetitive. It was assumed that semantically plausible distractors would produce unusually difficult items because of the competing distractor(s). Errors were expected to confuse students and thus create difficult items. Titles cueing items were expected to produce giveaways or easy items, while passage incoherence and violations in deletion patterns were expected to make items difficult.

Before these predictions about the relationships between identified flaws and statistically deviant items could be tested, certain calculations based on statistical data provided by the Rasch measurement model analysis had to be performed. Since these calculations and commentary thereon constitute the second phase of the total analysis of the cloze test forms, the full discussion of the procedures and implications related to these calculations will be withheld at this time in favor of a very brief summary. Questions arising in response to the following summary are referred to later sections of this chapter, Phase 2 and Phase 3.

The Rasch measurement model involves many statistical analyses, among them the analysis of the easiness of items (i.e., the proportion of students correctly responding to an item). Thus, if 100 students respond to an item with 75 answering correctly, the easiness of the item is .75. Since the Rasch model provided easiness data on every item on every passage on all 36 test forms, averaging the easiness of the items on a passage would give what was termed the passage easiness. A deviant item (for the purposes of this phase of the cloze analysis) was one whose easiness varied by a given amount from the passage easiness.

The Rasch model also provides response frequency data on every test item. That is, the Rasch model shows, given an item with five alternatives (i.e., the correct answer plus four distractors), how many times each of the alternatives was selected.

The predictive accuracy of the assumptions regarding the identified flaws was tested in terms both of item deviance and of distractor response frequency. That is, if an item had been identified as flawed, both its easiness score and the distribution of responses to its distractors were examined. Thus, it is conceivable that while an item flawed by a grammatically implausible distractor might not have a deviant easiness score, the flawed distractor might be dysfunctional or uncompetitive (i.e., a dysfunctional distractor is one which is not selected).

Observations. The testing of flawed items against item deviancy and response frequency data on a sample of 12 test forms (four from each of the three testing levels) produced the following observations. Generally speaking, the categorizations of flawed items proved to have little accuracy as predictors of item performance. Specifically, grammatically implausible distractors had almost no observable bearing on item deviance. Idioms, semantically plausible distractors, errors, and title cueing had some relationship to item deviance, but that relationship was relatively slight and often contrary to expectations. (Effects of passage incoherence and deletion pattern violations were not systematically observed at this time, but will be discussed during the description of Phase 3 of the total analysis.)

There was very little observable relationship between grammatically implausible distractors and item easiness scores. In other words, grammatically implausible distractors seemed to function no differently, no less adequately, than other distractors. The only exceptions involved a very few items in which distractors identified as grammatically implausible proved to be dysfunctional. These would include only those items which were difficult enough to produce a relatively high proportion of distractor selections. Given such an item, the one grammatically implausible distractor (i.e., out of four, usually) would have received virtually no attention. Put another way, grammatically implausible distractors, on items prompting much guessing, would be the distractors which even the guessers would refuse to select. Again, it must be stressed that the frequency of such items was very slight. Grammatically implausible distractors, then, have little utility as predictors of item deviancy.

Semantically plausible distractors, idioms, and errors were somewhat more useful as predictors of item deviance; they were often able to identify difficult items. To elaborate, a high proportion of distractors identified as semantically plausible proved, in fact, to be highly competitive with the correct answers. This was as expected. That flawed items categorized as idioms occasionally identified deviance is not unexpected, but it is somewhat surprising that idioms identified generally difficult items, rather than easy ones. Though the expectation was that idioms would make easy items, it transpired that, judging from the distribution of responses to distractors on such items, they apparently promoted much guessing. Idioms, then, did not behave exactly as predicted, but they did identify deviancy. Errors similarly produced difficult items, items involving more than usual degrees of guessing. This again is what one would expect; fortunately, there were very few items flawed by errors.

It was also observed that some titles which ostensibly cued items did indeed make such items generally less difficult. However, no pattern was observed in the effects on items or passages of titles which were misleading or irrelevant.

Though the categories of item flaws developed in the first step of this phase of the cloze test form analysis proved to have only slight utility in predicting item deviance, the experience of critically examining the cloze passages and compiling the categories has been very useful in several other ways. The critical examination of the test forms confirmed the need for a revision of the cloze rules; such a revision has occurred, and its utility has been discussed. The critical examination also pointed up the need for a systematic review of the entire cloze corpus, and this review, now underway, promises significant improvement of the cloze materials. Flawed items involving idioms, errors, and semantically plausible distractors have led both to the clarification and simplification of the cloze rules and to the improvement of existing cloze materials. Flaws involving grammatically implausible distractors, though identifying scant item deviancy, have also resulted in revisions of cloze rules and led to a review of the entire cloze corpus which will improve face validity of the materials. Further, the inconsistent effects of titles (combined with the difficulties implicit in attempting to control title-writing) have led to the decision to eliminate the title requirement in future applications of the cloze procedure. The item flaws have also made it clear that the most efficient way to identify item deviance is by inspection of the data resulting from administration of the test forms (see Phase 3). But the experience of examining test forms critically has augmented the acuity and sensitivity of the inspection of the test result data, especially by anticipating explanations of several typical kinds of item deviance.

Phase 2--Analysis of Readability Formula Utility

The second phase of the analysis of the cloze test forms involved the computation and categorization of passage easiness data for every passage on all 36 test forms. The first purpose of this analysis was to determine whether the test assembly procedures had produced test forms with certain similar (i.e., as nearly equivalent as possible) characteristics. Those characteristics were the range of passage difficulty per test form and the incremental pattern of difficulty among the passages on each test form at each testing level (i.e., Levels I, II, and III). The second purpose of this analysis was to determine the accuracy and utility of the Spache and Dale-Chall readability formulas as indicators of the relative difficulty of passages for students at given grade levels. This information was desired as a preliminary basis for guiding potential users of the cloze materials in the selection of passages.

The test assembly procedures were intended to produce 12 test forms at each of three testing levels, each form featuring passages of the same range of difficulty arranged in ascending order of difficulty. If test assembly procedures were successful, then easiness data (provided by the Rasch model) would show that the overall difficulty of a given test form did not vary dramatically from the overall difficulty of the other test forms at the same testing level, and that passages on all test forms were arranged so that each succeeding passage was more difficult than its predecessor.

The Rasch measurement model provides data on the easiness of every item on a given test form, arranged both within grade and across grades. That is, at a given test level, say, Level I, the Rasch model provides easiness data for grades 1, 2, and 3, and for grades 1 to 3 inclusive. For all 36 cloze test forms, passage easiness for each passage (i.e., the percentage of correct responses to an item, averaged by passage) was computed by grade and across grades. Table 8.6 in Appendix B contains passage easiness data for each test form, referenced to Test Development Notebook (TDN) identification numbers assigned to each passage.

The passage easiness data confirm that the desired pattern of passage difficulty on the test forms has generally been attained. For the most part, each succeeding passage on a given test form has a lower easiness than its predecessor. This pattern is consistent both across grades and within grades. The variations from this pattern are generally slight. Such variations, furthermore, are largely explainable by the expected overlap of readability scores. That is, the Dale-Chall readability formula produces a score which is then converted to a grade range. For example, a Dale-Chall score between 6.0 and 6.9 would, in conventional applications, indicate material for grades 7-8. For cloze purposes, however, such a score range was converted into four distinct readability levels (i.e., 13, 14, 15, 16). It is not surprising or unusual, then, that the readability scores on the cloze passages did not always predict actual passage difficulty with absolute accuracy.

Much of the overall slight variation from the desired pattern of increasing passage difficulty within a test form, then, was anticipated. But some of the variation is attributable to the accidental juxtaposition of somewhat deviant passages. For example, occasionally the second passage on a test form may be somewhat more difficult (on the basis of student performance) than most of the passages taken from its readability pool. Given that situation, and then given that the third passage on the test form may be an example of the opposite phenomenon (i.e., the passage selected may be somewhat easier than the other passages in its readability pool), such a juxtaposition of passages not surprisingly results in a variation from the desired pattern (i.e., the third passage on a test form is supposed to be more difficult than the second).

Generally speaking, then, the desired pattern of ascent in difficulty of the passages on each test form was attained. Thus, students taking the tests were confronted with similar tasks. If this had not been the case, if, say, some students had test forms with passages arranged in descending order of difficulty, then the tasks confronting the students, as well as the test-taking conditions or circumstances, would have varied greatly, thus resulting in widely differing anxiety levels among students and rendering questionable any illations derived from comparisons of test result data on different forms.

The basic similarity in difficulty of the test forms at each testing level is further verified by a comparison of the mean easiness[3] across grades on each test form at each testing level using the date in Table 8.1.

Table 8.1

Mean Passage Easiness Across Grades by Level

Level I		Level II		Level III	
Form	$\bar{X}$ Easiness	Form	$\bar{X}$ Easiness	Form	$\bar{X}$ Easiness
1	58.17	13	69.83	25	61.83
2	56.50	14	67.00	26	61.17
3	60.83	15	64.33	27	65.17
4	60.67	16	70.50	28	67.50
5	61.50	17	71.67	29	66.00
6	59.50	18	60.67	30	71.83
7	59.83	19	69.67	31	66.83
8	56.83	20	70.00	32	61.50
9	60.17	21	69.17	33	62.50
10	58.50	22	66.00	34	64.00
11	61.00	23	69.50	35	63.50
12	60.00	24	64.83	36	69.50

Note. The mean easiness ranges for Levels I, II, and III are 56.50-61.50, 60.67-71.67, and 61.17-71.83, respectively.

---

[3] Mean easiness provides an index to test form difficulty which is equivalent to the mean score. Mean easiness is the average *percentage* of correct responses, while mean score is the average *number* of correct responses.

As Table 8.1 shows, the variation in mean easiness scores (i.e., the average passage easiness of the six passages on a test form) per test form per level is slight. Again, grossly different mean easiness scores would have suggested some basic incomparability among the test forms, but the scores reveal no gross differences. Indeed, the fact that the range of mean easiness scores on the test forms at Level I is only 5.00 (i.e., low, 56.50; high, 61.50) suggests near equivalence among the test forms at that level. Such a narrow range of scores is as expected, since the passages for each test form were derived from a single domain and were systematically selected to include the same range of readability.

As expected, the range of mean easiness scores on the test forms at both Level II and Level III (i.e., 11.00 and 10.66, respectively) is broader than the range at Level I. These broader ranges seem intuitively reasonable, because one would expect to see, among students in grades 4 through 9, wider variations in literal comprehension.

Thus, both the conformity to expectations of the pattern of ascending passage difficulty on the test forms and the absence of any gross variations in mean easiness scores per test form per level seem to suggest that the test assembly procedures have succeeded in achieving the desired product: test forms similarly graduated and similar in difficulty. These surface similarities among the test forms lend credence to the proposition that they are equivalent measures of the same reading-related ability--literal comprehension.

Easiness averages for each passage on all 36 test forms are categorized (Table 8.7 in Appendix B) by passage readability levels both within grade and across grades. The easiness of every passage on every test form at a given testing level is compared to the easiness of all the other passages at the same readability level (determined by the Spache and Dale-Chall formulas). Thus, identification of passages which vary from expected performance is easy and convenient.

Just as the mean easiness of test forms at a given level spanned only a narrow range, so is the easiness range of passages at given readability levels similarly narrow. Specifically, at Level I the range of easiness of passages at a given readability level (i.e., identical to the readability level pools from which the passages were selected during test assembly) seldom exceeds .17 (e.g., the 12 passages from the pool representing readability levels 5 and 6 range in easiness from .43 to .59). Thus, any passage at Level I which broadens the easiness range at a given readability level beyond .17 would be considered deviant. The demarcation by which passages at Levels II and III were identified as deviant was based on an easiness range of .21. Given these ranges as touchstones for identifying deviant passages, of 216 total passages on the 36 test forms, only 11 were deviant (5%), 1 at Level I, 4 at Level II, and 6 at Level III. (An explanation of deviant passages is contained in the discussion of the third phase of the analysis.)

A more condensed schematization of the easiness of passages by readability levels is featured in Table 8.2. As this table illustrates, as the readability levels of passages increase, their easiness scores decrease. This pattern is generally maintained both within and across grades. An obvious and expected variation is that among students in higher grades the range of scores on passages at given readability levels is generally narrower and the upper extremity of scores is generally higher than among students at lower grades. An unexpected variation is that at Level III the upper extremity of scores on passages at readability levels 21 and 22 is higher among grade 8 than among grade 9 students. A possible explanation is that testing

Table 8.2

Distribution of Passage Easiness Scores by Readability Levels and by Grade, Level I

Passage sequence	Readability level	Grade 1										Grade 2										Grade 3										Grades 1-3									
		0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
1	1						7	2	2										4	7										6	5							1	7	3	
2	2					3	3	6	1									1	7	5										6	7							2	9	2	
3	3-4				6	4	2										1	6	5										2	7	3							9	3		
4	5-6			7	4	1									1	4	5	2									1	4	6	1						5	7				
5	7-8		1	11										1	2	4	4	1									3	4	5						1	8	3				
6	9-10		5	7										3	4	5										2	3	6	1						5	7					
	All levels		6	25	10	8	12	8	3					4	7	13	10	10	16	12						2	7	14	14	20	15				6	20	10	12	19	5	

Table 8.2 (Continued)

Distribution of Passage Easiness Scores by Readability Levels and by Grade, Level II

Passage sequence	Readability level	Grade 4										Grade 5										Grade 6										Grades 4-6									
		0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
1	5-6								4	8										11	1									7	5									11	1
2	7-8							4	4	4								1	5	5	1							1	3	6	2							1	5	6	
3	9-10						1	4	7									1	6	5									5	6	1							2	5	5	
4	11-12					3	4	4	1								2	4	6									6	3	3							4	4	4		
5	13-14				3	6	3									2	5	4	1								6	4	2							2	6	4			
6	15-16		1		5	6									1	7	4								1	2	7	2						1	1	7	3				
	All levels		1		8	15	8	12	16	12					1	9	11	10	18	21	2				1	2	13	13	13	22	8			1	1	9	13	11	14	22	1

Table 8.2 (Continued)

Distribution of Passage Easiness Socres by Readability Levels and by Grade, Level III

Passage sequence	Readability level	Grade 7										Grade 8									
		0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
1	11-12							2	4	5	1								5	7	
2	13-14						2	2	6	2								3	4	5	
3	15-16						5	5	2									4	7	1	
4	17-18				1	2	4	4	1							1	1	5	5		
5	19-20					7	4	3									6	8			
6	21-22			1	3	4	2							1		4	3	2			
	All levels			1	4	13	17	16	13	7	1			1		5	10	22	21	13	

Passage sequence	Readability level	Grade 9										Grades 7-9									
		0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
1	11-12								2	7	3								5	6	1
2	13-14							2	3	5	2							3	4	5	
3	15-16							3	6	2	1							6	5	1	
4	17-18					1	1	3	7							1	3	5	3		
5	19-20						8	5	[illegible]							1	10	3			
6	21-22			1	1	5	3							1		6	3				
	All levels			1	1	6	12	13	19	14	6			1		8	16	17	17	12	1

time for ninth-graders was limited because of variations in length of class periods.

Given that the readability formulas used to categorize passages by readability levels were employed in the full awareness and expectation that such formulas are only useful for making relatively coarse discriminations among passages, the easiness of passages within readability levels attests to a remarkable degree of accuracy in the determination of passage difficulty independent of test result data. Thus, the practical goal of the establishment of preliminary guidelines for the selection of passages by given readability levels has been met. Table 8.7 in Appendix B could serve as a guide to users of extant cloze materials in the selection of passages for the purpose of assembling test forms related to given readability levels.

The Rasch measurement model is a statistical technique for the calibration of passages by difficulty. It is a much finer and more sensitive method of calibration than readability formulas have been able to achieve, thus permitting an extremely accurate calibration of the cloze materials. Implementation of the Rasch model will greatly increase the utility of the cloze materials for each of their testing purposes (i.e., survey testing, achievement monitoring, and diagnostic testing).

Phase 3--Analysis of Item Deviance

The purpose of this phase of the analysis is fourfold. First, the analysis identifies the deviant items in a sampling of 12 of the 36 cloze test forms (4 at each testing level), discusses factors contributing to deviance, and draws tentative conclusions about the implications of item deviance for the continued development of the cloze materials. Secondly, the analysis discusses the 11 passages mentioned in Phase 2 as deviant, examines factors contributing to passage deviance, and suggests possible courses of action in response to passage deviance. Thirdly, the analysis makes observations and draws tentative conclusions about the relationship between deviance and item types (i.e., nouns, verbs, adjectives, adverbs). Finally, the analysis discusses the purposes and utility of the fit mean square both as an aid to passage calibration and as a method for identifying deviant items.

Deviant Items

Phase 1 of the analysis of the 36 cloze test forms had revealed the impracticality, as a method of identifying deviant items, of attempting to predict deviance based on flaws noted in the test forms prior to the examination of test result data. The experience of performing the Phase 1 analysis also suggested that an inspection of response frequency data, prior to or independent of consideration of easiness data, was similarly inefficient. Thus, experience showed that the most practical method of identifying deviant items was to examine easiness data.

An item was identified as deviant when its easiness score differed by a given amount from the average easiness of the items on the passage in which it occurred. Thus, to identify item deviance, Rasch easiness data were inspected and the easiness of each item on a test form was compared to the relevant passage easiness (computation of passage easiness was described in the discussion of Phase 2 of the analysis). Items on Level I test forms were identified as deviant when their easiness scores differed by ±.12 from the relevant passage easiness. A difference of ±.15 at Level II and ±.18 at Level III identified deviant items at those levels. These indices for the identification of deviant items varied by level because the range of item easiness scores per passage generally increased as the readability levels of the passages and the grade of the students increased.

After inspection of easiness data identified the deviant items on a given test form, each deviant item was examined thoroughly in order to determine, if possible, what factors (e.g., aspects of distractors, format, or passage) might have contributed to the deviance. If it were possible to generalize about characteristics of deviant items, clues might be discovered which would enhance the progressive improvement and refinement of the cloze materials. Thus, each deviant item was inspected within the context of its passage and in terms of the performance of its distractors. In other words, for each deviant item, both context and response frequency data were examined.

The results of this first step in the analysis of deviant items are presented in Table 8.8 in Appendix B. This table categorizes results based on analysis of 12 test forms, four from each testing level. Some of the test forms were randomly selected for analysis, and some were selected because the Phase 2 analysis had revealed that they contained deviant passages. Generally, the test forms analyzed are a fairly representative sample.

Table 8.8 in Appendix B is organized by testing level. Deviant items on the four forms analyzed at each level are identified by TDN passage identification numbers. Next, the table presents the following data and information: the easiness of the passage, the number of the item, the easiness of the item, the fit mean square of the item, and the part of speech of the item. Fit mean squares are provided only when they are high (i.e., above 3.70). Though they will not be discussed until the last step in Phase 3 of the analysis, the fit mean squares are provided here for the sake of future reference. Similarly, parts of speech (code numbers--1, 2,3,4--refer to nouns, verbs, adjectives, and adverbs, respectively) and item deviance will be discussed at a later time.

Finally, Table 8.8 in Appendix B offers interpretations of factors which seem to explain or contribute to item deviance. These interpretations are based on an analysis which considered both the nature of the passage containing the deviant item(s), and the behavior of the distractors. Thus, the interpretations refer both to context and to distractors. Since the interpretations of factors contributing to item deviance are, for the sake of economy, deliberately terse, each interpretive variation will now be explained, and illustrated where necessary.

Common association of words. Some words characteristically occur with great frequency in combination with certain other words. When words illustrative of such familiarly seen combinations are clozed, they typically result in unusually easy items. Thus, in the following phrases, when the underlined word is clozed, the resulting item will characteristically be easy for students at relevant grade levels: for the first *time*, ran all the way home from *school*, *brushed* her teeth.

Grammatically (syntactically) implausible distractor. Defined in the discussion of Phase 1 of the analysis, such distractors, though infrequently contributing to item deviance, do occasionally help to explain easy items. The following is an item whose deviance is partially explained by the grammatical implausibility of its distractors. "But everybody __________ you can buy . . ."; (a) swirls, (b) freezes, (c) pioneers, (d) *knows*, (e) gains.

Semantically plausible distractor. Also defined in Phase 1, distractors which can be substituted for deleted words without producing confusion or meaninglessness are obvious explanations for some difficult items. Interestingly, many of the deviant items which are explicable in terms of semantically plausible distractors seem to reveal that for many test takers, a distractor which is not semantically plausible in terms of the entire passage context will appear plausible, and thus function competitively, in a narrower or more restricted segment of context. Thus, the following item illustrates a distractor which, though not plausible in terms of the passage in its entirety, might indeed appear plausible if context were restricted to a single sentence: "I wish that you had bought us a ______ instead of another cow"; (a) star, (b) tower, (c) clock, (d) slab, (e) colt. Though the passage which included this sentence was about a family's frustrated desire for a clock, the sentence standing alone would seem to require distractor (e), colt, for completion.

Deleted word above level of passage. Readability levels of passages were based on Spache and Dale-Chall scores, which, among other factors, depend on percentages of hard words. These words are determined by word lists. The drawback of such word lists is that they do not distinguish the degrees of difficulty of words. Thus, occasionally passages may contain an extremely difficult word, the presence of which is obscured by the readability score. Given such a passage, if the very hard word is clozed, the resulting item will often, not surprisingly, be difficult. When deviant items seemed to be explicable in terms of the difficulty of the clozed word, such a resource as Harris and Jacobson's Basic Elementary Reading Vocabularies (1972) was consulted. Thus, in the following example, "the calf is weak and scrawny," the passage in question had a third grade readability score, but the clozed word, according to Harris and Jacobson, was a core 6 (sixth grade) word.

Idiom. Previously defined (Phase 1), idioms which seemed to contribute to item deviancy had the curious property of contributing to both hard and easy items. Apparently, on some items the distractors were so wildly implausible as to be largely uncompetitive, while on other items the strangeness of the distractors may have confused some students and led them to guess more than was expected. Example: "by a thousand or ______ of . . ."; (a) triangle, (b) wish, (c) more, (d) lilt, (e) sparrow.

Insufficient contextual clues. Occasionally, contexts from which given words are deleted do not obviously or explicitly cue the deleted words. In such cases items vary in difficulty depending on the adventitious behavior of distractors. In other words, in some cases where there is little or no explicit contextual cueing of an item, distractors may be plausible within a narrow segment of context, while in other similar cases it happens that distractors are not inordinately competitive. Examples of the first sort involve deviant items; examples of the second sort do not. Thus, in the following example, though the larger context clearly implies the correct answer, the absence of any explicit clue makes distractors (b) and (d) excessively competitive. "Everyone was bargaining ______ back and forth"; (a) snugly, (b) merrily, (c) loudly, (d) honestly, (e) painfully.

Title cues answer. A discussion of titles which cue correct answers was also presented in Phase 1. Most cases of titles cueing correct answers involve easy items. Example: title--Shopping at the Fish Market; item--"We came upon the fish market." However, there are exceptions to this generalization. Not every instance of a title containing a word which is subsequently deleted results in deviance. In other words, given titles which apparently cue items, such apparent cues do not necessarily alert a sufficient percentage of respondents to produce deviance.

Prior knowledge required. Deviance sometimes occurred when selection of the correct answer seemed related to some extent to the possession of prior factual knowledge. Thus, in a passage describing events and circumstances several years antecedent to the outbreak of the Revolutionary War, one of the items involved inferential knowledge based on the date of the signing of the Declaration of Independence (i.e., "six more years would pass"; unless students were familiar with historical dates, one of the distractors on this item, weekends, would be plausible as an indicator of time).

Colloquial expression. There are no doubt many examples of colloquial expressions or usage interspersed throughout the passages on the test forms. One of these examples seems clearly contributory to item deviance. The item in question, occurring in a passage of historical narrative, creates what may be called, for want of a simpler term, a breach in passage decorum. That is, the item, "the people of Boston had had it with British rule," represents a sudden and inappropriate introduction of a modern colloquialism into a passage which does not justify such usage.

Typographical error. Defined in Phase 1 under "error," typographical errors explain some difficult items.

Difficult sentence construction. Many difficult items seem to occur when words are deleted from complex or unusual sentence constructions. Example: "But being a calm and quiet young lady, she did not say anything, although the whole high school buzzed with rumors, guesses, reportedly authentic announcements on the part of students who had no right to be making announcements at all." The difficulty of the last deletion, guesses, seems a result of the appositional construction in which it functions.

Specialized word usage. Most words have more than one meaning or shade of meaning, and the variant meanings of some words differ substantially. Further, certain words have technical meanings associated with specific fields of endeavor. In either kind of instance, correct identification of such a meaning of a word depends to a certain extent on previously having experienced the word in its specialized use. Even such a relative commonplace as "the ragged pen," where pen is used synonymously for corral, resulted in a difficult item, partially, perhaps, because not everyone knows that a pen can be a corral.

Inexplicable. Deviant items for which no satisfactory explanation was discovered were thus categorized. Items so identified represent labored but unavailing examinations of passages and distractor behavior.

Distractor associated with context. There was one instance of a deviant item in which selection of a certain distractor was explained in terms of its association with some words in context. In the sentence, "Then everything in this world was dark and bitter for the minstrel of the gods," the distractor heavenly was frequently chosen. This was explained by the hypothesis that heavenly and gods have a typical association which, given gods in context, would make heavenly an attractive choice.

Table 8.3 summarizes by testing level the frequency of interpretations of item deviancy on the 12 test forms analyzed in Phase 3. A narrative consideration of Table 8.3 and Table 8.8 in Appendix B now follows.

Of the 31 deviant items on the four test forms analyzed at Level I, 11 items (36%) are easy and 20 (67%) are hard. Of the factors hypothesized as explaining or contributing to the deviancy of the easy items, common association of words represents 64% (7 items), titles cueing correct answers represent 27% (3 items), and idioms represent 9% (1 item). There is some overlap in the interpretation of the factors contributing to the deviance of the hard items at Level I. Of the factors hypothesized as explaining or contributing to the deviancy of the hard items, semantically plausible distractors repre-

Table 8.3

Frequency of Interpretations of Item Deviancy on Multiple-Choice Cloze Exercises

Interpretation	Level I	Level II	Level III
Common Association of Words	7	15	12
Syntactically Implausible Distractor			3
Semantically Plausible Distractor	9	11	5
Deleted Word Above Level of Passage	1	2	4
Idiom	3	4	2
Insufficient Contextual Clues	4	8	4
Title Cues Answer	3	2	1
Prior Knowledge Required	1	1	
Colloquial Expression	1		
Typographical Error		2	
Difficult Sentence Construction		5	16
Specialized Word Usage		2	3
Inexplicable	5	5	2
Distractor Associated with Context			1

sent 45% (9 items), insufficient contextual clues represent 20% (4 items), idioms represent 10% (2 items), and prior knowledge, deleted word above level of passage, and colloquial expression represent 5% (1 item) each. The deviancy of five of the difficult items (25%) was inexplicable.

Of the 52 deviant items on the four test forms analyzed at Level II, 18 items (35%) are easy and 34 items (65%) are hard. Of the factors hypothesized as explaining or contributing to the deviancy of the easy items, common association of words represents 83% (15 items), title cueing correct answer represents 11% (2 items), and idiom represents 6% (1 item). At Level II there is also some overlap in the interpretation of factors contributing to the deviance of hard items. Of the factors hypothesized as explaining or contributing to the deviancy of the hard items, semantically plausible distractors represent 32% (11 items), insufficient contextual clues represent 24% (8 items), difficult sentence constructions represent 15% (5 items), idioms represent 9% (3 items), typographical errors and specialized word usage represent 6% (2 items) each, and prior knowledge represents 3% (1 item). The deviancy of five of the difficult items (15%) was inexplicable.

Of the 44 deviant items on the four test forms analyzed at Level III, 14 items (32%) are easy and 30 items (68%) are hard. At Level III there is some overlap in the interpretation of factors contributing to the deviance of both easy and hard items. Of the factors hypothesized as explaining or contributing to the deviancy of the easy items, common association of words represents 83% (12 items), and title cueing correct answer and idiom represent 7% (1 item) each. Further, grammatically implausible distractors also seem to have a bearing on the deviance of 3 of the 12 items involving common association of words. Of the factors hypothesized as explaining or contributing to the deviancy of the hard items, difficult sentence constructions represent 53% (16 items), semantically plausible distractors represent 20% (6 items), insufficient contextual clues represent 13% (4 items), specialized word usage and deleted words above passage level represent 10% (3 items) each, and idiom represents 3% (1 item). The deviancy of two of the difficult items (6%) was inexplicable.

Generally speaking, by far the most frequent interpretation hypothesized in an attempt to explain deviant items which are easy is common association of words. Such a finding would seem intuitively reasonable; that is, since there clearly are many combinations of words which occur repeatedly in written discourse, it is not surprising to find that words in such combinations produce easy items when they are clozed. Indeed, it is desirable that a certain proportion of such familiar word combinations be represented in the cloze passages. To try to limit or restrict the frequency of occurrence of such word combinations in the cloze passages would lead to the production of highly biased and inordinately difficult or unusual testing materials.

Three factors seem most frequently to be associated with hard deviant items: semantically plausible distractors, insufficient contextual clues, and difficult sentence constructions. The influence and effect of the first two factors is to a large extent being attended to in the current review of the total corpus of cloze materials discussed in Phase 1. That is, all cloze passages and items are in the process of being corrected; semantically plausible distractors are being replaced and passages which contain items lacking sufficient contextual clues are being re-clozed.

That difficult or unusual sentence constructions are observed to have a recurrent relationship to difficult items is also intuitively logical. Sentence complexity does produce sentence difficulty. Again, since it is both necessary and desirable that the cloze materials be representative of the kind of reading materials that students actually encounter, it follows that a certain proportion of difficult, complex sentence constructions must occur in the cloze materials. Thus, an effort to avoid or edit out such sentence constructions would be misguided and counter-productive.

The analysis of item deviance shows, then, that the proportion of deviant items (approximately 20%) is not large. Current review of the cloze materials should reduce the proportion of deviant items both by correcting and by replacing potentially deviant items. And since deviance related to common word associations and difficult sentence constructions seems to attest to the relative freedom from bias in the procedures for selection and production of cloze materials, such deviance is desirable.

It should be noted further that certain aspects of the cloze materials related to deviance (specifically, common word associations concerning easy items and difficult or specialized words and difficult sentence constructions concerning hard items) suggest more about the weaknesses inherent in the Spache and Dale-Chall readability formulas than they do about the ostensible weaknesses of cloze materials. These two readability formulas, and perhaps especially the Dale-Chall formula, are particularly insensitive to the relative difficulty and easiness of words and to the relative difficulty of various kinds of sentence constructions. Because of the insensitivity of these readability formulas, then, passages within the same range of readability scores may not perform similarly because, despite their readability scores, they are in fact relatively easier or more difficult. In other words, the readability scores on such passages are very misleading. This problem, however, will be rectified with the implementation of Rasch passage calibration procedures which will rank passages in the cloze corpus by performance difficulty, thus providing a finer, more sensitive, and, hence, more practical and useful guide to the selection of cloze passages for test assembly purposes.

Deviant Passages

On the basis of inspection of the data (Table 8.7 in Appendix B) discussed in Phase 2 of the analysis, 11 passages (of the 216 total passages on the 36 cloze test forms) were identified as deviant. Table 8.4 lists these deviant passages by testing level, test form number, and TDN passage identification number. Further, the nature of the deviancy of each passage is also indicated (i.e., hard or easy). The following discussion attempts briefly to explain factors contributing to the deviancy of these passages, and concludes with several observations and recommendations.

Table 8.4

The Nature of Passage Deviancy on Multiple-Choice Cloze Exercises

Level	Form	Passage	Nature of Deviancy
I	3	04-08-01-01-01-020	Hard
II	15	04-09-01-01-05-038	Hard
	16	06-13-01-01-02-029	Easy
	18	04-07-01-01-03-013	Hard
	18	08-15-01-01-05-026	Hard
III	25	08-14-01-01-03-003	Hard
	26	07-13-01-01-01-010	Easy
	26	10-18-01-01-01-008	Hard
	30	08-16-01-01-05-018	Easy
	32	07-13-01-01-01-007	Hard
	33	10-22-01-01-01-029	Hard

Eight of these deviant passages were hard, and three were easy. Similar factors contributed to the difficulty of six of the eight hard passages. Thus, five of the six passages featured difficult sentence constructions, four manifested violations in deletion patterns, three contained flawed distractors, and two were marred by typographical errors.

Not only, then, were similar factors related to the difficulty of these six passages, but each of the six contained one or more remediable flaws. Thus, review should eliminate much of the deviant difficulty of these passages.

The other two difficult passages were irremediable. One featured obsolete or unfamiliar vocabulary (e.g., inkwells, and goose-quill pens) and the other was incoherent. These two passages could be profitably eliminated from the cloze corpus.

The three easy passages all had an uncommon degree of redundancy. Such redundancy is undetectable by readability formulas. Rasch calibration will more precisely calibrate these passages according to performance difficulty.

Eleven deviant passages from a total of 216 passages bespeaks a high degree of success in the development of procedures for constructing passages and test forms which perform consistently. Indeed, of the 11 deviant passages (three easy and eight hard), only two hard passages, one cluttered with archaisms and the other incoherent, seem to warrant elimination from the cloze corpus. No reapplication of cloze procedures could render these two passages acceptable. But correction of distractor flaws and adjustments of deletion patterns in the other six hard passages should improve them to the point where they no longer function deviantly. And further, the three easy passages will be precisely calibrated during the implementation of the Rasch model. Thus, given the rules revisions and the current review process discussed in Phase 1, cloze procedures seem to show considerable promise in assuring objectively reproducible test materials.

Item Deviance by Part of Speech

Deviance was analyzed in terms of the part of speech of items in order to determine whether items involving given parts of speech were characteristically deviant. If such a pattern were discovered, it might suggest that the part (or parts) of speech involved were not measuring the same aspect of reading ability as the other parts of speech. Furthermore, such a pattern would clearly indicate the need for closer and more detailed study of the behavior of items by part of speech.

Though the part of speech analysis has thus far been only a preliminary one, indications are that there is no discernible pattern of deviance associated with any of the parts of speech in question (i.e., nouns, verbs, adjectives, adverbs). Part of speech data presented in Table 8.9 and 8.10 in Appendix B contain no evidence which would implicate any of these parts of speech as characteristically deviant.

If these brief preliminary observations are confirmed by further study and analysis, then it may be concluded that cloze items involving nouns, verbs, adjectives, and adverbs are all relatively equally good and useful measures of literal comprehension.

Fit Mean Squares and Deviance

The Rasch measurement model is used to analyze test result data in order to determine whether the testing materials in question (i.e., items on test forms) are consistently measuring the same phenomenon or phenomena (in this instance, the reading-related ability called literal comprehension). Thus, the Rasch model assumes that all items on a test form are measuring the same characteristic or ability. In order to test this hypothesis, the model calculates a fit mean square for each item. When the fit mean square of an item exceeds a given point, the item is said to misfit. This implies that what the item is measuring seems to be at variance with what the other items are measuring. Since the cloze test forms are intended as a unidimensional measure of literal comprehension, it is assumed that fit mean squares on cloze items will help identify and rectify problem items.

The fit mean square analysis involved identifying every item (on the same 12 test forms analyzed earlier in Phase 3) with a fit mean square of 3.70 or higher[4] in order to determine the frequency of item misfit and to determine whether the misfit of items could be explained in ways similar to the hypotheses put forth (see Deviant Item Analysis) to explain items identified as deviant on the basis of variant easiness scores.

Results of the fit mean square analysis of 12 test forms are presented in Table 8.5. On the four test forms analyzed at Level I, there are 15 misfitting items, or 9% of the total items on the four forms. At Level II, there are 29 misfitting items, or 12% of the total items. At Level III there are 28 misfits, also 12% of the total items. On the face of it, then, there seems to be only a low proportion of item misfits on the 12 test forms. This would seem to indicate that the cloze testing materials conform to a unidimensional measurement model. If further, more sophisticated analyses of misfitting tend to confirm these preliminary findings, then the cloze testing materials will provide consistent and precise estimates of literal comprehension.

It should be noted that factors other than item quality can influence the fit mean square of an item. To wit, students who perform unusually and the element of chance can both contribute to high fit mean squares on given items. Since there are ways of accounting for the influence of such factors on apparently misfitting items, more detailed study and future experience

---

[4] Benjamin D. Wright and Ronald J. Mead, in personal consultation, recommended that a fit mean square of 3.70 or 4.00 be used as the determinant of item misfit.

Table 8.5

Analysis of Multiple-Choice Cloze Items With High Fit Mean Squares

Form	Item	Passage easiness	Item easiness	Fit mean square	Item deviancy	Part of speech
1	3	.75	.70	8.72		2
	34	.32	.22	6.71		2
	35	.32	.37	4.61		2
	37	.32	.35	4.60		1
3	15	.57	.38	5.77	X	3
	17	.57	.47	5.79		1
	22	.35	.35	12.88		1
8	17	.45	.34	3.95		2
	35	.41	.30	5.30		4
9	15	.56	.43	8.44	X	1
	26	.49	.47	3.75		1
	38	.40	.56	7.94	X	1
	39	.40	.29	5.90		1
	40	.40	.15	15.79	X	3
	41	.40	.37	7.39		1
14	1	.83	.84	8.98		2
	19	.72	.50	5.23	X	4
	35	.66	.45	6.98	X	4
	44	.53	.55	6.10		1
	47	.53	.36	3.94	X	4
	51	.46	.39	9.07		1
15	11	.74	.78	10.58		3
	18	.74	.59	5.93	X	1
	20	.74	.39	5.27	X	1
	25	.61	.29	4.59	X	1
	30	.61	.72	4.27		1
	52	.47	.28	4.86		2
	57	.47	.39	4.04		1
18	2	.84	.97	3.87		2
	4	.84	.55	8.08	X	1
	8	.84	.88	4.56		2
	12	.68	.47	10.04	X	2
	19	.68	.37	8.16	X	1
	35	.56	.26	4.73	X	2
	36	.56	.36	8.81	X	1

Table 8.5 (Continued)

Form	Item	Passage easiness	Item easiness	Fit mean square	Item deviancy	Part of speech
24	8	.80	.36	5.09	X	1
	16	.79	.85	5.34		3
	29	.74	.81	3.70		1
	37	.59	.14	5.34	X	3
	44	.50	.20	10.49	X	3
	48	.50	.45	6.04		2
	51	.47	.20	12.47	X	3
	56	.47	.29	4.69	X	3
	57	.47	.32	3.86	X	1
25	1	.74	.90	7.62		1
	2	.74	.95	3.70	X	2
	6	.74	.37	6.13	X	3
	9	.74	.45	5.00	X	1
	12	.61	.37	5.64	X	1
	14	.61	.45	5.14		1
	19	.61	.15	22.15	X	1
	31	.69	.25	7.79	X	1
	50	.48	.53	3.71		2
26	9	.74	.82	3.95		1
	10	.74	.44	8.22	X	3
	32	.41	.45	3.88		2
	35	.41	.23	3.77	X	1
	38	.41	.21	8.03	X	1
	47	.51	.64	4.43		1
30	10	.87	.92	15.37		1
	12	.77	.79	5.87		1
	13	.77	.60	6.82		2
	16	.77	.68	7.25		3
	17	.77	.67	14.64		2
	33	.71	.52	4.94	X	3
	52	.48	.56	4.03		4
36	11	.84	.58	4.61	X	4
	20	.84	.79	5.46		1
	22	.61	.41	8.86	X	1
	23	.61	.25	6.00	X	1
	43	.64	.47	4.08		4
	48	.64	.75	3.85		2

with the Rasch model may lower even further the percentage of misfitting items.

It is curious, at first glance, that there is so little overlap between misfitting items and easiness-deviant items. That is, many high fit mean square items do not appear deviant in terms of easiness variance, and many easiness-deviant items are not misfits. Thus, there is no necessary connection between misfitting items and items deviant in rather traditional terms. The explanation of this ostensibly strange lack of connection lies in the fact that the fit mean square is a function of the interaction of student ability and item difficulty. Thus, even given a flawed item which is very hard, students in given ability groups may perform according to statistical prediction. In such a case, the flawed item will not misfit. Similarly, on an apparently unflawed item students in given ability groups may not perform according to prediction. Such an item, therefore, would misfit. One conclusion which seems to follow from these observations is that since the fit mean square does not consistently identify items flawed in traditional terms, it should not be used for such identification purposes.

There is no necessary conflict here, however. Both the fit mean square analysis and the more traditional item deviancy analysis are essential to the refinement and calibration of the cloze materials. Flawed items must be rectified to the greatest extent possible, and misfitting items must be studied carefully in order that the measurement properties of the cloze be determined and consistently established.

At this preliminary stage of analysis, inspection of misfitting items seems to indicate a high degree of item consistency, or unidimensionality. But a further level of consistency, passage consistency, is desired for the cloze materials. A method of evaluating the extent of passage consistency (or unidimensionality), though developed, has yet to be implemented, and this method will rely heavily on the use of the fit mean square. Hence the importance of ongoing study into its uses.

Conclusions

The results of the three phases of the analysis of the 36 cloze test forms are generally very positive. Critical examination of the test forms independent of test result data has led to the revision of the cloze rules and to the initiation of a thorough review of all cloze materials. The rules revision is an essential step toward the eventual development of an algorithm which will permit maximum computerization and mechanization of cloze procedures. The revised rules also now make it possible for others to produce cloze passages and items with high consistency and quality. Further, the revised rules are a practical guide for the current review of cloze materials, a review which gives promise of assuring a final product which is free of flawed passages and items.

The analysis of the consistency of the passage graduation pattern per test form and of the difficulty (i.e., statistical easiness) of passages by readability level revealed a generally high degree of success. Variations in expected and desired patterns are largely attributable to the insensitivity of the readability formulas used in the initial scaling of passages. The findings of this phase of the analysis, then, suggest that the cloze test forms are remarkably consistent measures of literal comprehension. And the finer calibration of passages anticipated from the application of Rasch model procedures will eliminate the almost inevitable inconsistencies in passage scaling resulting from the use of readability formulas.

The analysis of deviant items is positive in several respects. First, the percentage of deviant items was low. Second, a large proportion of the flaws associated with deviant items will be corrected during the cloze review process. Finally, much of the item deviancy testifies to the relative absence of bias in the passage-selection and item-construction procedures. That is, the cloze materials seem an accurate sampling of reading materials actually encountered by students.

Preliminary analyses of passage deviancy, part of speech and deviancy, and fit mean squares are all encouraging. There were very few deviant passages, and most of those should be revised or adjusted by the review process and by the Rasch calibration of passages. That there was no discernible deviance pattern among any of the four parts of speech upon which cloze items were based suggests that these four parts of speech are nearly equally useful as measures of literal comprehension. Fit mean square analysis revealed a low proportion of misfitting items, thus implying a high degree of unidimensionality in the cloze test forms.

For an experimental effort, the administration of the 36 cloze test forms appears to have been highly successful in approaching the development of a consistent set of testing materials for literal comprehension. And the knowledge and experience gained through this experimental process promise greater success in future efforts.

Perhaps the most important foci for the further improvement of cloze procedures and materials are the effects of titles on cloze passages and more flexible passage formats (i.e., with greater length and more adequate context). On the other hand, further study of the function and utility of fit mean square analysis and implementation of the technique for identifying passage misfit are essential to the achievement of consistent unidimensionality among cloze testing materials and to the calibration of the cloze passages.

CHAPTER IX

RELIABILITY AND VALIDITY OF THE MULTIPLE-CHOICE CLOZE AND WH-ITEM TESTS

This chapter is concerned with what can be said to date about the reliability and validity of the literal comprehension measures. Presented here are the results of an initial exploration of the data from the May-June 1975 test administration first outlined in Chapters V and VI. These results represent the first stage of the analyses projected for the Multiple-Choice Cloze Exercises and other testing materials in the preliminary validation phase of the research (see Figure 6.1 in Chapter VI). This presentation is intended only as an early indication of the confidence that might be placed in the testing materials under development here. More detailed and definitive reports will be available later.

The discussion is organized as a research report that can be largely read and understood without detailed reference to the remainder of the text. Accordingly, presented first is a general overview of the study design, descriptive statistics for the major variables in the study, and a brief summary of the data analysis procedures.

The presentation of the results that follows begins with an examination of correlational data that reflect on the comparability of the cloze and wh-item test forms from two perspectives: (a) as measures of the same construct of literal comprehension and (b) as parallel test forms. This section of the results further analyzes the reliability and homogeneity of all test forms constructed for the research.

Next, the discussion is concerned with summarizing the results of the preliminary applications of the Rasch analysis program, provided by Wright and Mead (1975), to each of the test forms assembled in the cloze and wh-item formats. This is a very large set of results and a lengthy presentation is not attempted here. Instead, an attempt is made to summarize how well the total data set available on both the cloze and wh-item test formats fits the Rasch measurement model. Both the Rasch analyses and the more traditional analyses of internal consistency referred to in the previous paragraph reflect on the unidimensionality of the trait measured by these tests. In addition, these analyses provide a broad basis for evaluating the extent to which the multiple-choice cloze item form can be applied to samples of written discourse without seriously biasing the content of the testing materials.

The final section of the report presents the intercorrelations of scores on the literal comprehension measures and scores from the reading and language sections of the achievement test used in the study sample. Also included in the intercorrelation matrix are measures of verbal and non-verbal IQ and a measure of each student's test-wiseness. Together, these intercorrelations provide a rich context for exploring the construct validity of the cloze and wh-item tests, an exploration which begins only in brief with this report.[1]

Organization of the Study

The procedures for assembling the literal comprehension measures for this research were outlined in Chapter V as part of the presentation of potential applications of the cloze exercises. As reported in this discussion, a battery of multiple-choice cloze and wh-item test forms was assembled and administered to approximately 5,000 students distributed more or less evenly over grades 1-9. These tests were administered in May-June of 1975, in conjunction with the school district standardized testing program, thus making available additional test scores that could be used to explore the validity of the new tests of literal comprehension. The characteristics of the various tests that were used in the analyses are summarized below.

---

[1] More detailed analyses, to be conducted shortly, will examine the correlations of individual items on the California Achievement Test with multiple-choice cloze and wh-item scores.

Multiple-Choice Cloze Test

The multiple-choice cloze tests were assembled by systematically drawing passages from the Test Development Notebook (TDN) in the sampling design previously illustrated in Figure 5.5 in Chapter V. This design produced 3 sets of parallel test forms, with 12 forms in each of 3 successive test levels. A test form in Level I contained 6 passages, ranked by readability level, and 39 or 41 items. A test form in Levels II and III contained 6 passages, ranked by readability level, and 60 items. Test forms were randomly and evenly distributed across the student populations as follows: Level I--grades 1-3, Level II--grades 4-6, and Level III--grades 7-9.

Wh-Item Test

The wh-item tests were developed as an alternative measure of literal comprehension and assembled following a design that was virtually identical to that used in assembling the multiple-choice cloze tests. Like the cloze tests, each wh-test form contained 6 passages ordered in a fixed range of readability for each test level. However, the wh-item tests differed from the multiple-choice cloze tests in number of items per test form and in the item format. The wh-item tests each contained 30 multiple-choice test items, 5 for each passage. Each question was a verbatim transformation of a statement in the associated passage, based on one of the eight wh-item types (i.e., how, what [noun], what [verb], when, where, which, who, why).

Test-Wiseness Test

Because multiple-choice tests of reading comprehension are vulnerable to a form of test-wiseness referred to as "passage independence," a special test was constructed to measure some aspects of this characteristic. Referred to as the test-wiseness test, the design of these tests paralleled the cloze and wh-item test form designs. The wh-items in each of the three test levels were pooled and systematically assigned in units of 12 to each test form, such that no test items on this part of a given test were referenced to the passages on that test. Care was also taken to represent the passage difficulties and types of wh-items in a test level in an attempt to create parallel test forms. The relationship between scores on this test-wiseness measure and scores on the wh-item test provides some indication of the extent to which students' responses on the latter test are dependent on reading the associated test passages. This test also provides some indication of the extent to which this form of test-wiseness affects responses on the multiple-choice cloze test. However, it is probable that the cloze test is vulnerable to ot[illegible] forms of test-wiseness, a possibility that should be investigated directly.

Short Form Test of Academic Aptitude

The Short Form Test of Academic Aptitude (SFTAA) is a group-administered intelligence test that yields language and non-language IQ's. This test, administered by the school district along with the California Achievement Test, permitted study of the relationship between IQ and the literal comprehension tests across the study subsample.

California Achievement Test

The California Achievement Test (CAT), Form A, was administered to students in grades 1-8. The CAT has four levels as follows:

CAT Level	Grade(s)
I	1
II	2-3
III	4-6
IV	7-8

Each CAT Level provides scores on the following major skills:

1. Reading Vocabulary,
2. Reading Comprehension,
3. Language Mechanics, and
4. Language Usage.

In addition, the CAT provides reading and language subtest scores by test level as shown in Table 9.1.

Table 9.1

CAT Subtests by Test Level

Subtest	CAT Level I	II	III	IV
Sentence Picture Association	X			
Beginning Sounds	X			
Ending Sounds	X			
Letter Recognition	X			
Word Form	X			
Picture-Word Association	X			
Word Recognition	X	X		
Words in Context	X	X	X	X
Facts	X	X	X	X
Interpretation	X	X	X	X
Relationships			X	X
Generalizations		X	X	X
Inferences	X	X	X	X
Reading-General			X	X
Reading-Soc. Studies			X	X
Reading Science			X	X
Reading-Mathematics			X	X
Standard English	X	X	X	X
Sentence Structure			X	X
Sentence Parts & Functions			X	X
Transformations			X	X

All of these skill and subtest scores for the CAT were used to explore the meaning of the tests of literal comprehension developed for this study.

Analysis

The data set available on the foregoing test scores was organized for analysis by CAT level and by wh-item and multiple-choice cloze test level. There were four such subgroups: (a) students in grade 1 who took Level I on the CAT and Level I on the wh-item and multiple-choice cloze tests (N = 456); (b) students in grades 2 and 3 who took Level II on the CAT and Level I on the wh-item and multiple-choice cloze tests (N = 972); (c) students in grades 4, 5, and 6 who took Level III on the CAT and Level II on the wh-item and multiple-choice cloze tests (N = 1,699); and (d) students in grades 7 and 8 who took Level IV on the CAT and Level III on the wh-item and multiple-choice cloze tests (N = 594).[2]

To make this analysis possible, the raw scores for the wh-item and multiple-choice cloze test forms were converted to z scores based on the score distribution for each test in a test level. Subsequently, negative values were eliminated by applying a linear transformation to each set of obtained z scores. The resultant scores from any of the 12 wh-item and multiple-choice cloze tests in a test level were thereafter treated as having come from equivalent test forms and were combined as required for the analyses by CAT level. This approach to test equating, though somewhat unorthodox, is defensible on several grounds. The general shapes, means, and the standard deviations of the distributions of wh-item and multiple-choice cloze test scores were very similar from form to form in a test level (usually the average raw score difference from form to form was less than one-fourth of a standard deviation), the internal reliabilities of each form were consistently high, and the tests had been systematically assembled to be parallel in order and range of readability level.[3]

---

[2] The ninth grade is not included in the main analysis because the CAT was not given at this level.

The comparability of the cloze and wh-item tests as measures of a common construct of literal comprehension was examined by intercorrelating the various subscores and total scores on these tests by test level. The test-wiseness measure was included in these analyses. The reliability of all three types of tests was estimated by applying the Kuder-Richardson Formula 20 to the 108 available test forms.

The findings from the Rasch analysis of the wh-item and multiple-choice cloze tests are constrained to a summary of the fit of the available data to the Rasch measurement model. (The complete Rasch analysis on the 72 wh-item and cloze test forms is voluminous.[4] These analyses produced detailed item statistics, estimates of Rasch difficulty values for each item, ability estimates associated with each test score, fit mean squares for each item within specified ability groups, and bivariate plots of major item statistics.) The fit of the data to the model is determined by the mean and standard deviation of the fit mean square values within and across student ability groups. If the data fit the Rasch model, it can be concluded that the variable being measured is unidimensional. Of particular interest was the consistency of fit across forms and grade levels, which would reflect the stability of the trait when measured by virtually any systematically ordered test form that might be assembled from the multiple-choice cloze and wh-item passage and item pools.

---

[3] This same procedure was followed for the test-wiseness test forms, but here the justification for equivalence of test forms is less adequate.

[4] A description of a complete Rasch analysis is provided in Chapter VII of this proposal.

The construct validity of the multiple-choice cloze test is studied in the final analytical section of this chapter. The analysis is logical and correlational. Zero-order, Pearson product-moment correlations are used throughout this construct validation section. A priori specifications are made about the expected correlations between the multiple-choice cloze test and the wh-item test with all of the skill and subscores derived from the CAT. When expected correlations are not obtained, a detailed study is made of the possible causes for the unexpected results. These detailed studies focus directly on analyzing the item content of the CAT subtest in question in relation to the description of the construct of literal comprehension provided here.

To provide a basis for interpreting the results that follow, Table 9.2 displays the means and standard deviations of scores on the major variables under consideration. It will be observed that the multiple-choice cloze, wh-item, and test-wiseness tests were standardized (not normalized) within level. The major subtests on the CAT are normalized. The derived subscores on the CAT are raw scores.

There are several characteristics of this data set that must be taken into consideration when interpretations are made of the zero-order, Pearson product-moment correlations between variables. The construct validity sections of this chapter are based entirely on these correlations.

It is well-known that the reliability of a test as well as the shape of its score distribution affects its correlation with another test or measure. Reliability, as shown in the next section, does not contribute greatly to ambiguity in interpreting the results reported here. The wh-item and cloze tests meet the most conservative standards of reliability.

Table 9.2

Means and Standard Deviations of Scores on the Multiple-Choice Cloze, Wh-Item, Test-Wiseness, California Achievement, and Short Form Test of Academic Aptitude Tests

Test	CAT level							
	I (N=456)		II (N=972)		III (N=1,699)		IV (N=594)[a]	
	$\overline{X}$	SD	$\overline{X}$	SD	$\overline{X}$	SD	$\overline{X}$	SD
Multiple-Choice Cloze Test[b]	41.68	6.45	53.39	9.07	49.79	9.83	50.53	9.28
Wh-Item Test	40.64	8.12	53.87	7.63	49.69	9.92	50.37	9.28
Test-Wiseness Test	46.34	10.31	51.10	9.32	49.60	9.90	49.73	9.43
California Achievement Test								
Reading Vocabulary-ADSS[c]	316.11	34.52	370.14	46.78	430.87	64.66	533.31	82.75
Reading Comprehension-ADSS	298.39	48.29	389.00	65.03	461.01	68.74	545.50	81.41
Language Mechanics-ADSS	305.51	53.11	386.65	68.76	468.92	87.53	563.40	94.80
Language Usage-ADSS	327.08	57.48	393.22	67.97	459.25	75.02	541.79	77.83
Language IQ--SFTAA[d]	102.09	13.07	100.54	14.61	98.71	14.21	-	
Non-Language IQ--SFTAA	104.31	15.66	104.15	14.90	104.32	14.87	-	
Vocabulary Subscores[e]								
Sentence-Picture Association	9.87	.45	-		-		-	
Beginning Sounds	8.02	1.82	-		-		-	
Ending Sounds	8.97	1.43	-		-		-	
Letter Recognition	14.54	1.57	-		-		-	
Word Form	8.41	2.09	-		-		-	
Picture-Word Association	7.19	2.42	-		-		-	
Word Recognition	10.04	2.47	18.50	2.49	-		-	
Words in Context	5.45	3.90	14.42	4.84	24.83	8.83	27.13	8.36

Table 9.2 (Continued)

Test	CAT level							
	I (N=456)		II (N=972)		III (N=1,699)		IV[a] (N=594)	
	X	SD	X	SD	X	SD	X	SD
Comprehension Subscores:								
Facts	3.29	2.06	8.61	3.34	8.52	2.31	3.63	1.17
Interpretation	2.12	1.54	5.69	2.62	9.76	3.49	6.64	2.85
Relationships	-		-		1.52	1.16	3.84	1.66
Generalizations	-		3.87	1.82	3.97	1.94	4.70	2.30
Inferences	2.29	1.74	5.19	2.34	1.85	1.13	7.54	2.38
Reading-General	-		-		6.24	2.60	6.67	2.43
-Social Studies	-		-		4.59	1.71	5.95	2.44
-Science	-		-		5.08	2.32	4.95	2.40
-Mathematics	-		-		4.02	2.17	3.95	2.31
Language								
Standard English	11.70	3.88	14.70	5.17	14.42	3.52	15.67	3.77
Sentence Structure	-		-		3.35	1.22	2.68	1.59
Sentence Parts and Function	-		-		1.76	1.65	4.30	2.93
Transformations	-		-		2.51	1.29	4.04	1.30

[a] Grade 8 students only.

[b] Scores on the Multiple-Choice Cloze Test, Wh-Item Test and the measure of Test-Wiseness were standardized across grades 1, 2, and 3 (CAT Levels I and II); grades 4, 5, and 6 (CAT Level III); and grades 7, 8, and 9, (CAT Level IV plus grade 9) to have a mean of 50 and standard deviation of 10.

[c] Achievement Development Scale Scores (ADSS) were derived by CTB/McGraw-Hill from a single equal-interval score scale across all grades for use with all levels and forms of the CAT.

[d] IQ scores obtained from Short Form Test of Academic Aptitude given with CAT.

[e] All California Achievement Test Subscores are raw scores.

The test-wiseness test is sufficiently reliable for the purpose used here. A few of the CAT subtests should be suspect because of length; they are, at Level III, Relationships (4 items), Inferences (4 items), Sentence Structure (5 items) and Transformations (5 items), and, at Level IV, Facts (5 items). The potential low reliability of these subtests is noted in appropriate tables. The subscores for the wh-item tests, which are based on small numbers of each of the wh-item types in the test, and the deletion type subscores for adjectives and adverbs on the cloze test, which are typically based on very few items, undoubtedly contribute to reduced correlations among the subscores and total scores for these tests. However, these relationships were as expected, and they do not seem to have contributed to interpretive problems, particularly when the high degree of internal consistency of the wh-item and cloze tests is shown later with other methods.

A minor problem for interpretation of the correlational results derives from the test score distributions of the cloze and wh-item tests which are generally skewed in the negative direction. This skewness is the result of generally easy tests which interact with grade level in this case to produce more extreme negative skewness at the upper grades of a test level and also across the grades of the total study sample. The general result of this skewness is a frequent lack of homoscedasticity at the upper score levels in the bivariate plots of the test score distributions in the present study. This effect appears to have produced attenuation in certain of the correlations at certain levels of the study sample which is predictable as follows.

In grades 1, 2, and 3, and particulary in grade 1, the wh-item and cloze tests were considerably more difficult for students than in other grades. As a result there was no artificial ceiling on the tests, and students' scores were spread more evenly over the score ranges for the two tests. Given the high reliability of the Level I wh-item and cloze tests for grades 1, 2, and 3, the correlations of the Level I cloze and wh-item tests with each other and with the CAT Level I and II skill and subtest scores should be close to the "true" correlations among these variables.

In CAT Levels III and IV of the study sample, the degree of negative skewness on the wh-item and cloze tests is more marked, with the Level IV data being the most affected. At these test levels, the attenuating effects of skewness on the correlations involving the wh-item and cloze tests should be noticeable, particularly at Level IV. The overall effect of the skewed distributions of the scores on the two measures of literal comprehension constructed for the study is to reduce the validity coefficients between the cloze and wh-item tests and between these tests and the various CAT scores in the upper levels of the study sample.

At Level IV, the validity coefficients are further reduced because of an undetected error in the construction of the original CAT data tape.[5] Grade 7 data were misplaced, so that the Level IV correlational data are based solely on grade 8 students. This results in a severe restriction in the range of student ability at CAT Level IV with an associated reduction in variability. This reduction in variability is the major cause for the low, zero-order correlations at CAT Level IV.

Test Comparability and Reliability

The correlations among subscores and total scores within and between the three types of tests specially constructed for this research are presented in Tables 9.3, 9.4, and 9.5. The three total test scores are

---

[5] Prepared by CTB/McGraw-Hill.

Table 9.3

Intercorrelations of Multiple-Choice Cloze, Wh-Item, and Test-Wiseness Tests: Total Scores and Subscores
Level I: Grades 1, 2, and 3 (N = 1,445)

Scores	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
MCC Test															
1 Noun	-														
2 Verb	.85	-													
3 Adjective	.70	.69	-												
4 Adverb	.49	.45	.35	-											
5 Total	.96	.94	.78	.53	-										
Wh-item Test															
6 How	.63	.62	.47	.29	.64	-									
7 What (N-P)	.64	.62	.46	.32	.65	.58	-								
8 What (V)	.64	.61	.47	.32	.65	.60	.59	-							
9 When	.56	.55	.41	.28	.57	.50	.46	.49	-						
10 Where	.59	.58	.43	.28	.61	.54	.56	.55	.50	-					
11 Which	.65	.61	.47	.33	.65	.63	.60	.61	.51	.57	-				
12 Who	.66	.64	.46	.30	.67	.63	.64	.62	.52	.59	.65	-			
13 Why	.59	.59	.44	.30	.61	.52	.55	.54	.47	.50	.55	.55	-		
14 Total	.79	.77	.58	.39	.81	.79	.80	.80	.69	.77	.82	.84	.73	-	
Test-Wiseness Test															
15 Total	.32	.30	.24	.19	.32	.23	.22	.24	.21	.25	.28	.28	.23	.31	-
Mean	49.7	49.6	49.7	49.7	49.7	49.6	49.7	49.6	49.6	49.5	49.6	49.6	49.5	49.5	49.6
S.D.	9.9	10.0	9.6	7.6	10.0	10.0	9.9	9.9	9.9	9.9	10.0	9.9	10.0	9.9	9.9

Note. Scores were standardized across grades 1, 2, and 3 to have a mean of 50 and S.D. of 10.

Table 9.4

Intercorrelations of Multiple-Choice Cloze, Wh-Item, and Test-Wiseness Tests: Total Scores and Sub cores
Level II; Grades 4, 5, and 6 (N = 1,728)

Scores	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
MCC Test															
1 Noun	-														
2 Verb	.87	-													
3 Adjective	.79	.77	-												
4 Adverb	.65	.64	.58	-											
5 Total	.97	.94	.87	.72	-										
Wh-Item Test															
6 How	.51	.52	.46	.40	.53	-									
7 What(N-P)	.49	.49	.44	.34	.51	.45	-								
8 What(V)	.48	.49	.44	.37	.51	.43	.44	-							
9 When	.48	.47	.42	.35	.49	.45	.41	.43	-						
10 Where	.51	.48	.43	.37	.51	.47	.42	.41	.44	-					
11 Which	.52	.52	.46	.38	.54	.49	.47	.45	.47	.48	-				
12 Who	.52	.50	.45	.38	.53	.45	.43	.42	.43	.44	.50	-			
13 Why	.53	.52	.47	.40	.55	.47	.45	.44	.43	.48	.47	.48	-		
14 Total	.70	.70	.62	.52	.73	.73	.70	.70	.70	.72	.75	.71	.73	-	
Test-Wiseness Test															
15 Total	.28	.28	.27	.22	.30	.22	.23	.22	.23	.22	.24	.25	.25	.33	-
Mean	49.7	49.7	49.6	49.6	49.6	49.6	49.5	49.5	49.5	49.5	49.5	49.6	49.5	49.6	49.5
S D	9.9	9.9	9.9	9.9	9.9	10.0	9.9	9.9	10.0	10.0	9.9	10.0	10.0	10.0	9.9

Note. Scores were standardized across grades 4, 5, and 6 to have a mean of 50 and S D of 10.

Table 9.5

Intercorrelations of Multiple-Choice Cloze, Wh-Item, and Test-Wiseness Tests: Total Scores and Subscores
Level III; Grade 8 (N = 594)[a]

Scores	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
M-C Test															
1 Noun	-														
2 Verb	.82	-													
3 Adjective	.78	.75	-												
4 Adverb	.58	.56	.53	-											
5 Total	.96	.91	.87	.66	-										
Wh-Item Test															
6 How	.35	.32	.31	.26	.36	-									
7 What (N-P)	.39	.38	.35	.25	.41	.49	-								
8 What (V)	.34	.36	.33	.27	.38	.44	.42	-							
9 When	.38	.38	.36	.29	.41	.43	.43	.43	-						
10 Where	.40	.37	.34	.26	.41	.44	.46	.43	.41	-					
11 Which	.39	.37	.34	.26	.40	.47	.47	.44	.47	.46	-				
12 Who	.40	.42	.34	.27	.42	.45	.44	.44	.47	.51	.51	-			
13 Why	.45	.39	.39	.26	.45	.41	.50	.45	.45	.49	.46	.46	-		
14 Total	.54	.52	.48	.37	.56	.71	.73	.69	.70	.73	.73	.73	.73	-	
Test-Wiseness Test															
15 Total	.23	.19	.22	.12	.23	.10	.16	.09	.15	.17	.09	.11	.11	.17	-
Mean	50.48	50.39	50.63	50.02	50.53	50.54	49.93	50.25	50.03	50.11	50.06	50.67	49.93	50.37	49.73
S D	9.30	9.48	9.58	9.81	9.28	9.36	9.79	9.42	9.57	9.65	9.52	8.81	9.52	9.28	9.43

Note. Scores were standardized across grades 7, 8, and 9 to have a mean of 50 and S D of 10.

[a] Grades 7 and 9 are not included.

based on the wh-item test, the multiple-choice cloze test, and the test-wiseness test. The subscores for the multiple-choice cloze test are each based on one of the four grammatical parts of speech deleted in a passage. For example, each test form has a subscore for the sum of all correct noun responses. The subscores for the wh-test are based on the various types of wh-items included in the test. The test-wiseness test has only a total score.

Tables 9.6 and 9.7 present, respectively, the means and standard deviations and the Kuder-Richardson Formula 20 reliability coefficients for each of the 36 forms of the three types of tests. All Tables (9.3-9.7) are organized by test level on the wh-item and cloze tests (Level I for grades 1-3; Level II for grades 4-6; and Level III for grades 7-9).[6]

Test Comparability

The issue of test comparability has two important facets of concern here. The first is the purely technical one of determining whether the various test forms constructed for the research were reasonably parallel and thus could be combined in the analyses by subsample level.

The other facet of test comparability has to do with determining the extent to which the wh-item and multiple-choice cloze tests measure the same thing. These tests were originally constructed as alternate methods of measuring the same construct--literal comprehension. Tables 9.3-9.5 are relevant to the latter concern, though these data are not a basis for strong inferences about the parallelism of the constructs measured by the wh-item and multiple-choice cloze tests. The intercorrelations in Tables 9.3-9.5

---

[6] Table 9.5 is for test Level III but the data are for grade 8 only. Not all of the correlational analyses were completed in the 7-9 sample at the time of this writing.

are presented in three sections. For example, Table 9.3 presents the intercorrelations of the deletion type subscores for the Multiple-Choice Cloze Exercises in the upper left-hand triangular section of the matrix; the wh-item subscores for the Wh-item Test are in the lower-right triangle; and the intercorrelations of the subscores of the two tests are presented in the lower-left rectangular section of the table. The correlations of the subscores of both tests with the total test scores are given in rows 5 and 14 of the matrix. The correlations of all scores with the test-wiseness score are given in the last row of the table. These data support the following observations in the Level I sample:

1. The deletion type subscores for the multiple-choice cloze test forms are moderately to highly intercorrelated, except for the adverb score which tends to have low correlations with the other subscores.

2. The noun score correlates highest with all other deletion type subscores and with the total test. (The correlation of nouns with the total test is particularly high, but this is due to the fact that the test is largely composed of noun responses at Level I).

3. Although the wh-subscores are each based on only a few items (the average equals 3.5), these subscores are consistently and moderately interrelated.

4. The wh-item subscores generally correlate at about the same level with the total wh-test score.

5. The correlations of noun or verb scores with the wh-item subscores are generally as high as the intercorrelations of the wh-item subscores.

6. The correlation between the total scores on both the multiple-choice cloze and wh-item tests is .81, a level at which either test is fairly predictable from the other. (If these tests were two norm-referenced measures of reading comprehension, this correlation would be interpreted as a very acceptable validity coefficient.)

7. The test-wiseness score has a significant but very low correlation with the other test scores.

The foregoing observations hold for the other levels of the multiple-choice cloze and wh-item tests given in Tables 9.4 and 9.5, with a few important qualifications. The correlations of adjective and adverb scores with subtest scores and total test scores on the multiple-choice cloze test improve across grades 4-8, due apparently to changes in the composition of passages at these more advanced reading levels. (The mean and variance of adverb and adjective subscores increase, but the total test is still predominantly noun and verb.) The correlations between subscores and total scores on the wh-item and multiple-choice cloze tests tend to decrease across grade levels, with this decrease being most marked in the grade 8 sample (r between total test scores is .56, down from .81 in the Level I subsample).

These data lead to the conclusion that the two types of literal comprehension tests assembled for the research are generally consistent within themselves across the study sample. The pattern of the intercorrelations within and between test types is further consistent with the interpretation that the subscores in either test contribute to a single factor and this factor is common to both tests. However, the commonality that appears evident between the two tests is substantially reduced in the upper grade samples, an effect that is apparently attributable to the shapes of the score distributions at these levels and reduction in range of talent.

The other factor of test comparability of interest here may be evaluated in part by reference to Table 9.6, which presents the means and standard deviations for each test form in the wh-item and multiple-choice cloze formats. In addition to these data, the proportions of correct responses for each item in each test form were arrayed and visually inspected as were the raw score distributions on each test form. Inspections of the

Table 9.6

Means and Standard Deviations for the Multiple-Choice Cloze and Wh-Item Tests

	Multiple Choice Cloze				Wh-Item Test			
Level	Form	N	$\bar{X}$	S.D.	Form	N	$\bar{X}$	S.D.
I (Grades 1, 2, 3)	1	128	21.03	10.55	37	127	18.80	7.15
	2	126	20.26	10.49	38	124	19.52	7.32
	3	130	21.51	10.59	39	126	19.57	7.29
	4	124	22.44	11.34	40	121	19.02	7.30
	5	126	23.06	11.24	41	119	18.43	7.72
	6	126	19.71	9.24	42	122	19.17	7.00
	7	127	21.47	11.22	43	124	19.05	7.28
	8	127	18.84	10.40	44	124	19.20	7.53
	9	129	21.98	11.31	45	131	19.10	7.69
	10	127	20.47	10.43	46	123	19.19	7.47
	11	120	23.39	11.25	47	121	18.65	7.18
	12	123	22.67	11.41	48	121	20.12	7.69
II (Grades 4, 5, 6)	13	147	41.46	11.45	49	147	22.74	5.57
	14	151	40.01	14.11	50	153	21.95	5.46
	15	153	38.73	12.51	51	148	22.72	4.85
	16	152	40.99	11.62	52	152	22.74	5.66
	17	146	42.18	12.60	53	145	23.52	5.46
	18	151	36.35	11.03	54	144	23.19	5.45
	19	152	41.80	13.48	55	145	22.96	5.00
	20	148	42.00	12.08	56	149	22.60	5.96
	21	152	41.39	11.37	57	147	20.76	6.11
	22	152	39.63	13.57	58	148	22.19	5.86
	23	148	41.72	12.99	59	157	23.76	5.51
	24	149	39.01	13.32	60	145	21.87	5.73
III (Grades 7, 8, 9)	25	167	36.60	12.53	61	163	23.81	5.54
	26	164	36.44	11.69	62	162	23.89	7.01
	27	160	38.86	14.33	63	164	24.25	5.79
	28	161	40.47	12.82	64	161	23.89	4.83
	29	158	39.17	11.52	65	165	23.53	4.75
	30	165	42.54	13.35	66	166	21.20	6.20
	31	158	39.46	12.45	67	154	24.88	4.85
	32	163	37.07	12.01	68	163	22.40	5.65
	33	166	37.38	11.98	69	164	24.02	4.99
	34	159	38.08	13.60	70	156	22.01	5.31
	35	163	37.82	13.18	71	163	23.16	5.94
	36	165	41.82	12.56	72	154	22.03	6.90

means and standard deviations of each set of test forms shows that the difference between any pair of test-form means is relatively small (e.g., for the Level I multiple-choice cloze test forms, the largest difference between means is less than one-third of the average standard deviation; for the Level I wh-item test, the average difference between means is less than one-fourth of the average standard deviation). These data, taken together with the examination of the score distributions on each wh-item and multiple-choice cloze test form, indicate that the various forms of each test type in a test level were fairly comparable. Given the other properties of each test form (e.g., progressive ordering of passages by readability level), it then appears that the various test forms of each type in a level would result in fairly comparable scaling of students at each level of the wh-item and multiple-choice cloze tests.

Test Reliability

The reliability of a test is of interest because it estimates the amount of random error contained in a test score. The validity coefficients reported here for the wh-item, multiple-choice cloze, and test-wiseness tests cannot be interpreted satisfactorily without an appropriate estimate of measurement error. The reliability statistic selected for this purpose was the Kuder-Richardson Formula 20 or KR-20 (Kuder and Richardson, 1937), which is a special case of the Hoyt (1941) or Cronbach (1951) coefficients of equivalence when test items are scored dichotomously. As used here, KR-20 reflects directly on several other properties of the cloze and wh-item tests. The formula provides an estimate of the homogeneity of the items in the test or the proportion of test variance attributable to the first general factor in the test. In addition, the KR-20 formula provides a good estimate of the short-term stability of the test. And, since the

KR-20 is available here on a large number of test forms systematically assembled from the TDN, the average or median value of the KR-20 is also a basis for judging the reliability of the process of test generation.

Accordingly, the KR-20 also reflects on the validity of the wh-item and multiple-choice cloze tests. The description of the construct indicates the test is a measure of a homogeneous trait that should be stable over short periods of time. Inspection of the KR-20 coefficients in Table 9.7 for the multiple-choice cloze and wh-item tests indicates that reliability expectations for these tests were confirmed at a very high level. The median KR-20 value for the 12 cloze test forms at each test level was .96; for the wh-item test, the KR-20 ranged from .91 to .93. These data provide further support for the conclusion that each of the two measures of literal comprehension is a highly reliable estimate of a single homogeneous trait. The tests assembled to measure the trait should thus scale individuals similarly over short periods of time, and it appears that similar tests for measuring this trait can be repeatedly assembled.[7]

A final point of comment concerns the reliability of the test-wiseness test forms assembled for each test level. Across all test levels, 4 of these test forms had low reliabilities (Forms 40, 45, 47, and 53), but given the shortness of these test forms and their unusual composition, the KR-20's are surprisingly high. These results are similar to Tuinman's (1973) for a similar test he referred to as a measure of passage independence.

Rasch Analysis

As explained in Chapter VII, complete analyses using the Rasch measure-

---

[7] This inference is supported directly and indirectly by the data reported here and by the consistency of the results reported in the next section of this report.

Table 9.7

Kuder-Richardson Formula 20 Reliability Coefficients for the Multiple-Choice Cloze, Wh-Item, and Test-Wiseness Tests

	Multiple-Choice Cloze Test					Wh-Item Test					Test-Wiseness Measure			
Level	Form	N	I	KR-20	SE	Form	N	I	KR-20	SE	Form	N	I	KR-20
I (Grades 1,2,3)	1	128	41	.94	1.73	37	127	30	.92	2.02	37	127	12	.76
	2	126	41	.95	1.64	38	124	30	.94	1.79	38	124	12	.68
	3	130	41	.96	1.46	39	126	30	.90	2.30	39	126	12	.68
	4	124	41	.96	1.46	40	121	30	.90	2.31	40	121	12	.47
	5	126	41	.95	1.73	41	119	30	.91	2.32	41	119	12	.67
	6	126	39	.95	1.57	42	122	30	.91	2.10	42	122	12	.79
	7	127	41	.96	1.45	43	124	30	.93	1.92	43	124	12	.68
	8	127	39	.96	1.51	44	124	30	.90	2.38	44	124	12	.51
	9	129	41	.97	1.33	45	131	30	.90	2.43	45	131	12	.26
	10	127	41	.96	1.49	46	123	30	.91	2.24	46	123	12	.78
	11	120	41	.96	1.43	47	121	30	.94	1.76	47	121	12	.13
	12	123	41	.96	1.54	48	121	30	.92	2.18	48	121	12	.76
	Median			.96	1.49				.91	2.21				.68
II (Grades 4,5,6)	13	147	60	.97	1.98	49	147	30	.93	1.47	49	147	12	.50
	14	152	60	.96	2.82	50	153	30	.93	1.44	50	152	12	.75
	15	153	60	.96	2.50	51	148	30	.90	1.53	51	148	12	.68
	16	152	60	.96	2.32	52	152	30	.86	2.11	52	152	12	.74
	17	146	60	.97	2.18	53	145	30	.93	1.44	53	145	12	.29
	18	151	60	.94	2.69	54	144	30	.92	1.54	54	144	12	.76
	19	152	60	.97	2.33	55	145	30	.85	1.94	55	145	12	.46
	20	148	60	.95	2.69	56	149	30	.95	1.33	56	149	12	.73
	21	152	60	.95	2.53	57	147	30	.94	1.50	57	147	12	.71
	22	152	60	.97	2.35	58	148	30	.91	1.76	58	148	12	.58
	23	148	60	.97	2.25	59	157	30	.94	1.35	59	157	12	.66
	24	149	60	.95	2.97	60	147	30	.93	1.51	60	145	12	.76
	Median			.96	2.35				.93					.70

Table 9.7 (Continued)

	Multiple-Choice Cloze Test					Wh-Item Test					Test-Wiseness Measure			
Level	Form	N	I	KR-20	SE	Form	N	I	KR-20	SE	Form	N	I	KR-20
III(Grades 7,8,9)	25	167	60	.96	2.51	61	163	30	.91	1.66	61	163	12	.55
	26	164	60	.95	2.61	62	162	30	.94	1.71	62	162	12	.70
	27	160	60	.96	2.87	63	164	30	.96	1.16	63	164	12	.77
	28	161	60	.97	2.22	64	161	30	.89	1.60	64	161	12	.78
	29	158	60	.96	2.30	65	165	30	.89	1.57	65	165	12	.59
	30	165	60	.97	2.31	66	166	30	.94	1.52	66	166	12	.77
	31	158	60	.95	2.78	67	154	30	.96	0.97	67	154	12	.75
	32	163	60	.96	2.40	68	163	30	.90	1.78	68	163	12	.79
	33	166	60	.95	2.67	69	164	30	.96	0.99	69	164	12	.77
	34	159	60	.95	3.03	70	156	30	.93	1.40	70	156	12	.70
	35	163	60	.97	2.28	71	163	30	.95	1.32	71	163	12	.75
	36	165	60	.97	2.17	72	154	30	.95	1.56	72	154	12	.72
	Median			.96	2.40				.94	1.54				.75
Overall	Median			.96					.92					.72
	Mean			.96					.92					.65
	Range			.94-.97					.85-.96					.13-.79

Note. N = number of subjects.
I = number of items.

ment model were calculated on all forms of the multiple-choice cloze and wh-item tests. These analyses provide evidence for answering the following questions: Do the multiple-choice cloze and wh-item tests measure one trait? Is the measurement of this trait consistent across grade levels? As noted previously, both tests were designed to measure one trait, literal comprehension. The Rasch analysis provides a further test of this assumption as well as additional evidence on the generalizability of the cloze item form to levels of written discourse.

> The Rasch model specifies a particular simple relationship between person ability, item difficulty, and the probability of observing a correct response. The implications of this specification are that:
>
> 1) the variable measured is unidimensional;
>
> 2) there are no strong relationships among the persons or items other than those specified by the model so that responses of persons to items are stochastically independent given their parameters in the model;
>
> 3) items and persons do not differ substantially with respect to other possible response factors not represented in the model such as item discrimination, person sensitivity, guessing or indifference. (Wright and Mead, 1975, p. 2).

If the data analyzed for the present study fit the Rasch measurement model, then these three conditions of the model must have been satisfied in the available response data. More specifically, if the data on the multiple-choice cloze and wh-item tests fit the Rasch measurement model, then it can be concluded that the variable being measured by each test is unidimensional.

Table 9.8 displays the mean and standard deviation of the fit mean square statistics for all items in each form of the multiple-choice cloze test. This average fit mean square is calculated from each item fit mean square, which is the appropriate statistic for testing the fit of each item to the Rasch model. These mean fit mean square statistics have expected

Table 9.8

Mean and Standard Deviation of the Fit Mean Square Statistics for Each Form of the Multiple-Choice Cloze Test

Form	N	Mean of fit mean square	S D of fit mean square
1	126	1.60	1.69
2	124	1.32	0.83
3	130	1.74	2.19
4	124	2.00	2.59
5	124	1.46	0.88
6	126	1.56	1.79
7	125	1.97	1.82
8	125	1.43	1.04
9	127	2.52	2.88
10	126	2.09	2.02
11	119	1.91	1.63[a]
12	121	1.92	2.28
13	147	2.03	2.58
14	152	1.74	1.77
15	152	2.15	1.67
16	151	1.63	1.53
17	146	1.30	1.08
18	151	2.08	1.94
19	153	1.83	1.71
20	149	2.16	2.95
21	152	2.28	2.84
22	152	1.79	1.80
23	148	1.62	1.17
24	147	1.99	2.01
25	166	2.51	3.08
26	163	1.98	1.53
27	160	2.11	2.04
28	161	2.34	2.07
29	158	1.80	1.58
30	162	2.39	2.79
31	158	2.60	3.20
32	163	2.03	1.57
33	166	1.56	1.53
34	156	1.81	1.33
35	162	2.50	2.49
36	165	1.81	1.61

[a] Denotes expected S D of .35. All other forms have expected S D of .28.

values of 1.0. The standard deviations of these mean fit mean square statistics have expected values of the square root of 2 over the degrees of freedom for the number of score groups. For 6 and 5 scores groups, the expected standard deviation equals .28 and .35, respectively.

With reference to Table 9.8, it is possible to determine the best- and worst-fitting test form. The best-fitting form will have a mean and standard deviation close to the expected values of 1.0 and .35 (Or .28, for 5 instead of 6 score groups. The worst-fitting form will have values furthest from these expected values. Based on these criteria, the best-fitting multiple-choice cloze test is Form 2; the worst fitting is Form 31. There are no multiple-choice test forms with statistics that deviate radically from the expected values. However, test calibrations with rather high (i.e., above 2.0) mean fit mean squares should be studied in detail to determine the cause of misfit. Due to the fact that the forms of the test range from grade 1 to grade 9, it can be concluded that the trait measured by the multiple-choice cloze test is unidimensional and stable across these grades.

A more detailed analysis of the fit of the multiple-choice cloze data to the Rasch model is provided in Table 9.9. This analysis is more sensitive than the previous analysis because the fit statistics are calculated within score groups. These score groups increase in ability from the first to the sixth group. Score ranges for the score groups are determined by the program so as to make the N of each group as equal as possible, based upon a predetermined minimum group size.

The fit statistics in Table 9.9 are mean and standard deviations of $\underline{z}^2$ statistics for testing the fit of each score group. Under the assumption that the multiple-choice cloze data fit the Rasch model, the mean $\underline{z}^2$ statistics have expected values of 1.0 and standard deviations of 1.4. This anal-

Table 9.9

Means and Standard Deviations of the $\underline{z}^2$ Statistics for Testing the Fit of Each Item in Each Score Group for Multiple-Choice Cloze Test Forms

	Student sub-groups																	
	First group			Second group			Third group			Fourth group			Fifth group			Sixth group		
Form	N	$\overline{X}$	SD	N	$\overline{X}$	SD	N	$\overline{X}$	SD	N	$\overline{X}$	SD	N	$\overline{X}$	SD	N	$\overline{X}$	SD
1	20	1.9	3.2	22	2.0	3.8	23	1.0	1.1	20	1.2	1.4	20	1.5	2.2	21	2.0	4.8
2	20	1.2	1.6	22	1.7	2.3	20	1.1	1.0	19	0.9	1.1	20	1.3	1.8	23	1.7	2.5
3	24	2.9	6.8	23	2.3	3.6	23	1.1	1.5	22	1.2	1.4	23	1.3	1.4	15	1.7	3.4
4	21	2.9	6.8	20	2.2	6.4	19	2.4	5.1	21	1.1	1.3	21	1.4	2.2	22	2.0	2.3
5	20	1.4	1.4	18	1.4	2.1	21	1.1	1.2	20	2.0	3.1	22	1.4	1.6	23	1.4	1.6
6	20	1.7	2.7	23	1.1	1.4	22	2.0	4.6	22	0.9	1.4	21	1.5	2.1	18	2.1	4.1
7	20	2.8	7.3	23	2.3	3.2	21	1.6	1.9	21	1.7	1.8	20	1.7	2.0	20	1.7	2.8
8	23	1.6	3.2	18	1.1	1.8	21	1.4	2.0	20	1.2	1.8	22	1.8	2.2	21	1.5	3.5
9	19	2.5	3.5	22	4.5	8.0	22	3.1	5.9	22	1.7	2.5	25	1.9	2.6	17	1.4	3.5
10	19	2.9	3.9	19	2.8	6.2	24	1.1	1.6	19	1.4	1.8	21	2.1	2.9	24	2.3	3.0
11	22	2.1	2.3	24	2.0	3.9	24	1.3	1.8	25	1.5	1.5	24	2.7	4.5	0	0.0	0.0
12	19	2.6	5.3	20	2.4	6.5	22	1.4	1.7	20	1.3	1.5	18	2.1	2.1	22	1.8	3.1
13	22	4.8	12.3	23	1.4	1.9	26	1.2	2.2	23	1.0	1.1	25	1.7	2.6	28	2.0	5.3
14	25	3.2	6.4	25	1.1	2.3	26	1.4	1.6	25	1.2	1.3	28	1.9	3.7	23	1.8	3.4
15	23	3.1	4.1	25	1.6	2.1	27	1.3	1.7	24	1.7	2.3	27	2.5	2.9	26	2.7	5.0
16	25	2.5	4.6	24	1.2	1.7	24	1.2	1.5	26	1.3	1.7	25	1.8	4.1	27	1.7	2.7
17	25	1.8	2.5	23	0.8	0.9	24	0.8	0.9	21	1.4	2.5	26	1.3	1.6	27	1.9	5.0
18	26	4.6	8.2	27	1.7	1.8	25	1.0	1.2	24	1.7	2.2	25	1.6	2.7	24	1.9	3.1
19	25	3.8	6.5	26	1.7	2.7	26	1.6	2.0	22	1.2	1.5	23	1.3	1.6	31	1.6	3.7
20	24	4.5	10.3	26	1.8	2.1	24	1.4	1.7	22	1.4	1.7	25	1.2	1.0	28	2.7	8.9
21	23	3.4	7.9	26	3.4	7.6	25	1.3	1.9	23	1.7	1.9	25	1.7	1.9	30	1.9	3.4
22	25	4.0	8.7	27	1.5	2.1	25	0.9	1.1	27	1.0	1.4	25	2.2	3.1	23	1.2	1.3
23	24	3.6	5.9	23	1.3	1.7	24	0.9	0.9	27	1.6	1.8	27	1.4	2.1	23	1.0	1.3
24	24	4.0	10.1	25	2.0	3.0	21	1.3	1.9	26	1.7	3.2	25	1.3	1.4	26	1.6	2.4

Table 9.9 (Continued)

	Student sub-groups																	
	First group			Second group			Third group			Fourth group			Fifth group			Sixth group		
Form	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD
25	27	5.2	14.9	29	1.9	3.0	27	1.7	2.5	23	2.1	4.9	26	1.5	2.0	34	2.7	3.3
26	28	2.9	6.3	30	2.3	2.8	25	1.4	1.6	27	2.1	2.6	25	1.3	1.4	28	1.9	2.2
27	25	3.6	7.2	27	1.7	3.3	25	1.3	1.4	24	1.8	2.2	25	1.7	2.7	34	2.5	4.2
28	27	3.5	5.9	26	2.9	3.0	26	1.1	1.5	28	2.2	2.8	30	2.2	3.1	24	2.2	6.7
29	26	2.6	4.0	27	2.0	4.9	25	1.2	1.4	29	1.9	2.7	26	1.1	3.5	25	2.1	3.5
30	27	2.4	3.6	28	1.6	2.2	29	1.8	2.1	26	2.0	2.9	29	4.7	12.9	23	1.8	4.3
31	25	3.7	14.4	27	3.1	4.0	25	1.2	2.1	26	2.1	2.6	23	2.4	2.7	32	3.1	5.4
32	27	3.3	4.9	29	1.8	2.9	28	1.4	1.7	29	2.1	3.2	29	1.8	2.0	21	1.9	2.5
33	29	3.8	7.4	27	0.9	1.4	27	1.1	1.6	26	1.0	1.2	29	1.3	1.7	28	1.3	2.0
34	26	2.5	3.6	26	2.1	3.5	28	1.2	1.6	28	1.4	1.6	25	1.8	2.0	23	1.8	3.1
35	25	3.1	7.8	27	3.6	6.8	26	1.8	2.4	27	2.4	2.8	27	2.0	2.2	30	2.3	3.6
36	28	3.3	5.7	25	1.7	2.0	30	1.3	1.5	31	1.5	2.1	28	1.6	3.0	23	1.4	4.0

ysis pin-points the location of "misfit" with the Rasch model within a particular ability group. It must be emphasized that "misfit" is relative in this and subsequent analyses because it is primarily due to student ability and not test items. (In actual practice, when one is calibrating items on a test, one removes the students, not the items, that are causing the "misfit.")

Again, the best-fitting multiple-choice test is Form number 2. The greatest deviation from expectation is found in the high ability group with a mean of 1.7 and a standard deviation of 2.5. (Form 2 had practically a normal distribution of scores.) The multiple-choice cloze test with the poorest fit is again Form number 31. Form 31, with the ability groups so specified, does not fit well in 3 of the 6 groups. In addition, the standard deviations of the mean fit statistics are far removed from expected values. These results are consistent with the previous analysis.

The results in Tables 9.8 and 9.9 support the conclusion that the trait, namely literal comprehension, measured by the multiple-choice cloze test is unidimensional and stable from grades 1 to 9.

A final point should be noted concerning the fit, within ability groups, of the multiple-choice cloze test. It is commonly observed in Rasch analyses of student test data that "misfit" is constrained to low and high ability groups. In the case of the multiple-choice cloze, the low ability groups are causing the most problem. As previously noted, in calibration work, several low-ability students would be deleted and the Rasch analysis rerun on the same form. This second run would display a better fit of the data to the Rasch model.

Table 9.10 presents the mean and standard deviation of the fit mean square statistics for all items in each form of the wh-item test. When compared to the same statistics for the multiple-choice cloze test, the

Table 9.10

Mean and Standard Deviation of the Fit Mean Square Statistics For Each Form of the Wh-Item Test

Form	N	Mean of fit mean square	S D of fit mean square
37	126	2.06	2.09
38	121	1.74	1.33
39	122	1.94	1.26
40	120	2.15	1.88
41	114	2.38	2.25[a]
42	120	2.34	2.39
43	123	1.84	1.26
44	123	1.85	1.85
45	129	2.27	2.12
46	123	4.96	13.23
47	118	2.19	2.22[a]
48	106	1.42	.89[a]
49	142	1.36	1.07
50	149	1.61	1.30
51	147	1.79	2.31
52	148	1.73	1.58
53	141	1.46	1.15
54	137	2.12	1.98
55	142	1.49	1.44
56	137	1.30	1.19
57	140	1.10	0.69
58	144	1.63	1.21
59	136	1.56	1.17
60	142	1.46	1.83
61	163	1.69	3.03
62	161	4.72	18.06
63	153	2.14	2.46
64	154	1.56	0.97
65	165	1.42	1.79
66	161	1.59	1.96
67	145	1.47	1.49
68	160	1.62	1.27
69	160	1.49	1.08
70	146	1.19	0.83
71	162	3.05	10.86
72	145	1.38	1.44

[a] Denotes expected S D of .35. All other forms have expected S D of .28.

values for the wh-item test are closer to expected values. This finding was expected because more stringent passage controls were used in the development of the wh-item test than the multiple-choice cloze test. This consistency is reflected in all of the values reported in Table 9.10. Note that the worst-fitting forms are 46 and 62.

A more detailed fit analysis of the wh-item is provided in Table 9.11. Again the worst-fitting wh-item tests are Forms 46 and 62. Other than these extreme forms, there is a very consistent pattern of effects in Table 9.11. The low ability group seems to account for nearly all of the extremes in misfit. Put another way, if some of the low ability students in these analyses were removed and the wh-item test data recalibrated using the Rasch model, the fit of the data to the model would be even more consistent. Generally, from these results it can be concluded that the trait measured by the wh-item test, which is also hypothesized to be literal comprehension, is also unidimensional and stable from grades 1 through 9.

The conclusions drawn from the Rasch analyses support the conclusions drawn in the previous section on test comparability and reliability. Generally, the items on the various wh-item and cloze test forms contribute to the measurement of a single, homogeneous trait.

Table 9.11

Means and Standard Deviations of the $\underline{z}^2$ Statistics for Testing the Fit of Each Item in Each Score Group for Wh-Item Test Forms

	Student Sub-group																	
	First group			Second group			Third group			Fourth group			Fifth group			Sixth group		
Form	N	$\overline{X}$	SD	N	$\overline{X}$	SD	N	$\overline{X}$	SD	N	$\overline{X}$	SD	N	$\overline{X}$	SD	N	$\overline{X}$	SD
37	21	2.3	3.2	21	3.1	6.3	21	1.6	1.5	21	1.8	1.7	24	1.8	2.1	18	1.8	3.4
38	21	2.6	2.7	24	2.2	4.5	22	1.8	2.0	24	1.3	2.1	23	2.0	3.5	7	0.5	1.1
39	19	3.2	5.5	19	2.4	3.7	20	2.6	2.5	23	1.0	1.1	17	1.2	2.0	24	1.3	2.2
40	21	5.7	8.7	19	2.0	2.9	21	1.6	1.5	20	1.3	1.5	20	1.1	1.3	19	1.3	2.7
41	22	4.6	8.9	22	1.5	2.4	22	3.0	3.1	24	1.5	1.6	24	1.4	2.0	0	0.0	0.0
42	20	4.2	5.8	19	3.6	10.9	22	2.2	2.5	20	1.8	2.0	22	1.3	1.5	17	1.1	2.1
43	20	3.5	6.4	21	1.5	2.1	21	1.7	1.6	17	1.6	1.4	20	1.3	1.3	24	1.5	1.8
44	21	4.6	8.8	20	1.1	1.3	18	1.8	2.0	17	1.0	1.1	24	1.2	1.4	23	1.4	2.4
45	22	4.5	6.9	22	2.9	7.3	25	2.1	2.3	18	1.2	1.7	19	1.8	2.5	23	1.3	1.7
46	20	6.0	10.1	19	15.7	66.8	18	2.0	2.4	24	2.6	2.7	19	2.0	2.0	23	1.6	3.4
47	22	2.1	2.9	25	2.7	5.6	24	1.8	2.9	23	1.5	1.4	24	2.9	5.8	0	0.0	0.0
48	21	1.7	2.3	20	1.6	2.5	23	1.1	1.3	22	1.5	1.6	20	1.2	1.5	0	0.0	0.0
49	21	1.5	2.0	25	1.1	1.3	19	0.8	0.7	26	1.6	2.3	30	1.4	3.0	21	1.8	3.3
50	26	1.4	1.6	21	1.0	1.2	26	1.3	1.7	27	1.7	2.1	22	1.0	1.2	27	3.3	5.2
51	26	3.7	6.7	20	0.8	1.0	27	1.3	1.6	30	1.3	1.7	22	2.1	4.7	22	1.5	3.7
52	27	3.6	6.5	26	1.3	1.6	27	1.6	1.7	24	1.7	1.3	29	1.6	2.6	15	0.6	1.2
53	24	2.1	3.6	22	0.7	0.8	26	1.3	1.4	16	1.6	3.0	18	1.0	1.3	35	2.1	2.9
54	22	4.3	7.7	24	1.3	1.9	22	1.5	1.7	16	1.0	0.9	30	2.1	2.3	23	2.4	6.3
55	24	2.2	3.6	26	1.1	1.3	15	1.1	1.5	18	1.1	1.2	23	1.0	1.2	36	2.5	6.5
56	24	2.7	4.8	25	0.9	1.1	26	1.4	1.5	25	0.8	0.8	27	0.7	0.9	10	1.2	2.3
57	22	1.2	1.3	24	0.8	1.0	23	1.0	1.3	27	0.8	1.0	20	1.1	1.3	24	1.8	3.2
58	21	3.0	4.4	25	1.2	1.6	19	1.2	1.7	19	1.3	2.0	23	1.5	2.3	37	1.6	1.6
59	21	2.0	3.0	24	1.2	1.6	26	1.4	2.0	22	1.5	2.2	26	2.2	3.4	17	1.2	1.5
60	23	3.3	8.5	25	0.8	1.0	26	0.9	1.1	22	1.0	1.1	24	1.2	1.4	22	1.7	3.4

Table 9.11 (Continued)

	Student sub-group																	
	First group			Second group			Third group			Fourth group			Fifth group			Sixth group		
Form	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD
61	27	3.0	7.6	29	2.3	6.7	33	1.8	2.8	23	0.9	0.7	31	1.7	2.4	20	0.4	1.3
62	26	20.4	99.2	25	2.8	8.8	22	1.3	1.3	19	1.0	1.0	20	1.0	1.4	49	1.9	5.0
63	25	3.7	8.8	25	1.5	1.9	29	1.3	1.3	32	2.6	3.2	22	1.1	1.2	20	2.6	9.0
64	22	2.4	3.2	28	1.0	1.0	23	1.5	1.9	25	1.9	2.7	21	1.2	2.0	35	1.4	1.7
65	32	2.5	8.5	19	1.0	1.5	27	1.0	1.5	15	0.6	0.6	27	1.4	1.8	45	2.1	2.9
66	27	2.2	3.3	26	1.3	1.6	31	1.2	1.7	22	1.6	2.8	26	1.6	2.8	29	1.7	5.4
67	28	2.0	2.9	22	0.8	1.1	19	0.7	0.7	27	1.4	3.0	31	2.0	3.2	18	2.0	5.2
68	25	2.7	4.7	26	1.3	1.4	25	1.1	1.2	30	1.0	1.2	30	1.8	2.4	24	1.8	2.1
69	28	2.5	5.4	24	1.6	2.3	32	1.6	2.2	16	1.0	1.1	22	0.9	1.1	38	1.3	1.6
70	25	1.8	3.0	19	1.1	1.3	19	1.4	1.2	26	0.9	2.0	16	0.7	0.7	41	1.3	2.0
71	27	12.3	62.4	32	1.2	1.7	33	0.9	1.0	22	0.8	1.0	17	1.6	4.2	31	1.6	2.5
72	25	1.9	3.0	26	1.2	1.5	24	1.1	1.6	29	1.7	2.1	24	1.2	1.9	17	1.2	3.8

Construct Validation

The principal focus of construct validation as discussed in Chapter VI, is the explication of the network of interrelated concepts that define the trait in question and the conditions and interpretations that surround its measurement. One accepted approach to construct validation is to examine the convergence and divergence between the principal measure in question and other measures that are indicators of the same construct or can be discriminated from the construct. Another approach is to examine the convergence of two alternate measures of the same construct, obtained by dissimilar methods, with other measures that are indicators of the same construct or can be discriminated from the construct (Cronbach, 1971).

In the present analysis, the first approach consisted of analyzing expected levels of correlation between multiple-choice cloze scores and various other test scores available in the study--the wh-item test, the CAT skill scores and subtest scores, language and non-language IQ test scores, and test-wiseness scores. In the second approach, the convergence of the multiple-choice cloze and wh-item tests as measures of literal comprehension was evaluated in a simultaneous comparison of the values of the correlations of both of these tests with the other test criteria available in the study.

In the first approach, convergent validity is evaluated by the general consistency with which predictions are confirmed in terms of relative levels of correlation between the multiple-choice cloze test and other measures in the study considered to be similar indicators of the construct of literal comprehension or related or unrelated indicators of other constructs. Some measures are expected to correlate relatively highly with the multiple-choice cloze test, others relatively moderately, and still others are predicted to have a low correlation. In the second approach, convergent validity is determined by reference to the size of the differences in the absolute values of the correlations of the multiple-choice cloze and wh-item tests with the other test criteria available in the study. These absolute differences are expected to be small, consistent from measure to measure, and generalizable across levels of the study population.

In general, the predictions indicate that the multiple-choice cloze test is relatively highly related to the wh-item test, unrelated to the measure of test-wiseness, moderately related to the measure of language IQ, and related in varying degrees to the CAT, depending on the "apparent meaning" of a skill or subtest score denoted by the test label. Both the labels and specific item content of these CAT skill and subtest scores present some difficulties in interpretation, inevitably leading to ambiguity concerning whether a given test is a similar measure of the construct or has some other relationship with literal comprehension. Where these ambiguities arise in the present analysis, disconfirming predictions, an attempt is made to resolve the problem by examination of the test items in question. Such an analysis, however, is potentially fraught with the usual problems of all post hoc analyses. That is, some explanation of the disconfirming event can usually be found, and for this reason, a post hoc analysis must be taken as exploratory or hypothetical.

Convergent and Divergent Validity of the Multiple-Choice Cloze

It was assumed that predictions about the correlations between the multiple-choice cloze test and the other test criteria available in the study could be made on the basis of the construct definition of literal comprehension and the information at hand defining the content of the test criteria. In order to understand the rationale for these predictions, it is instructive to review briefly the construct definition and the availability of information that constitutes an adequate definition of each of the various potential indicators or non-indicators of it.

The construct: the multiple-choice cloze test. The construct, literal comprehension, and its relationship with the multiple-choice cloze test are stated in detail in Chapter IV of this proposal. In brief, literal comprehension is the apprehension of "the grammatical and semantic relations which obtain within and among the sentences of the discourse" (Katz and Fodor, 1967, p. 172). The multiple-choice cloze test accesses these grammatical and semantic relations by systematically deleting nouns, verbs, and modifiers[1] from a segment of written discourse, and then placing the deleted words in sets of responses where the distractors are all grammatically plausible but semantically implausible. It is hypothesized that students will have no difficulty in selecting the only word which is grammatically and semantically plausible if they can apprehend "the grammatical and semantic relations which obtain within and among the sentences of the discourse." The distractors in other words, do not function as traditional distractors--do not, in fact, "distract"--until the syntactic and semantic complexity of the discourse exceeds the students' psycholinguistic competence. The test is designed, therefore, to discriminate between a specifiable set of interactions--called literal comprehension--between student and text, and another specifiable set of interactions between student and text called no comprehension. The test is designed, that is, to measure literal comprehension or no comprehension and nothing else. The interactions between student and test--the extensiveness of the processing of the grammatical and semantic relations in the text--are carefully controlled by the type and rate of deletion and the distractor selection procedure. The item type is hypothesized to access only literal meaning; it should access no nuances of meaning and no other semantic interrelationships than those clearly signaled in the grammatical and semantic relations of the text.

---

[1] Only nouns and verbs are deleted in grade 1 and 2 materials.

The construct: the wh-item test. The wh-item test is designed to access the same grammatical and semantic relations of a given text. Like the multiple-choice cloze, the wh-item test is considered an indicator of the construct, literal comprehension. The wh-item accesses the grammatical and semantic relations of a text by deleting immediate constituents in clauses of the sentences in the text, replacing them with the appropriate wh-words (who, what, which, where, when, how, or why), and then transforming the clauses into questions.[2]

The wh-item, then, is the traditional question type teachers use to direct student attention to salient features of the text, and correct answers to such questions are usually considered evidence of literal comprehension. The primary difference between such traditional questions and the wh-item test is the systematic way in which the wh-item is written. Such systematization makes it possible to specify and control, to a greater degree than possible with traditional test questions, the interactions between the features of the text and the psycholinguistic competence of the student.

Since the wh-item test has some claim to a specifiable relationship with the construct definition, it is considered the preferred or least ambiguous criterion measure for the multiple-choice cloze in the analysis that follows. The primary difficulty with using the wh-item as a criterion measure is that it is subject to a form of test-wiseness discussed in Chapter II. In brief, it is possible for a student with minimal syntactic competence to locate the correct answers to wh-items in the text without understanding what the question or the text means. However, as will be seen throughout the course of the analyses, there is no reason to suspect that the wh-item test was actually subject to this form of test-wiseness in the study samples.

---

[2] The wh-item test is described in more detail in Appendix A.

The construct: other criterion measures. The test-wiseness measure used in this analysis was reviewed in the introduction to this chapter and is not to be confused with the form of test-wiseness just discussed. The test-wiseness measure used in the analysis is a preliminary effort to determine the passage dependence of the wh-items. Passage dependency is also crucial to the construct, literal comprehension. That is, the test items must access only the grammatical and semantic relations which obtain within and among the sentences of the discourse.

The language and non-language IQ scores reported in the following analysis come from the Short Form Test of Academic Aptitude. In the development of the construct through Chapters I to IV, it was hypothesized that language IQ scores should only correlate moderately with scores from tests of literal comprehension, since the literal level of comprehension requires little of the inferential and related reasoning processes so characteristic of measures of verbal intelligence. Non-language IQ scores would seem to have little or no relation to the construct of literal comprehension and should thus correlate to a lesser degree with the multiple-choice cloze than language IQ.

The problems that exist in specifying the relationship between the construct definition of literal comprehension and each of the skill and subscores of the CAT have already been noted. Given the lack of an explicit statement defining the psycholinguistic meaning of each CAT skill and subscore used in the analysis, it was necessary to define the relationship of these skills with the construct by recourse to the test score labels and the meager skill or subscore descriptions given in the CAT Test Coordinator's Handbook (1970).

Cat skill scores: predicted correlational levels. The predicted correlational levels between multiple-choice cloze scores on the one hand, and the wh-item test, CAT skills, and language and non-language IQ scores on the other, are presented in Table 9.12. These predictions are based upon the relative degree to which it is expected the different measures will converge upon or diverge from the construct, literal comprehension. The predictions for the CAT scores have been based on the labels or brief descriptions attached to a skill score or subscore, and the consistency of the application of the labels is assumed.

Table 9.12

Expected Levels of Correlation of Multiple-Choice Cloze Scores with Wh-Item Test Scores, California Achievement Test Skill Scores, Test-Wiseness Scores, and Language and Non-Language IQ Scores

Lowest (.00-.29)	Medial (.30-.54)	Highest (.55+)
Test-Wiseness	Language Usage	Wh-Item Test Scores
Language Mechanics	Language IQ	Reading Vocabulary
Non-Language IQ		Reading Comprehension

The highest predicted levels of correlation as evidenced by Table 9.12, are between the multiple-choice cloze scores and the wh-item test scores, the CAT reading vocabulary, and the CAT comprehension scores. The crucial correlation is, of course, between the multiple-choice cloze and the wh-item test, the preferred criterion measure. The two experimental tests must correlate highly with each other. A strong prediction is also made about the correlations between the multiple-choice cloze scores and scores from the test-wiseness measure, for reasons already cited. The remaining predictions are less strong because the relationship between the construct and the remaining measures is not so clear. Language mechanics and nonverbal IQ scores are expected to have low correlations with multiple-choice cloze scores while language usage and language IQ scores should fall in the medial range.

Also, implicit in Table 9.12 is the assumption that the multiple-choice cloze test will behave consistently across grade levels as long as passages are properly matched in readability with the psycholinguistic competence of the students. Such consistent behavior is crucial to the possibility of using the same item type to measure literal comprehension, regardless of the content of the test passages or the reading ability of the student. Hence, the predictions in Table 9.12 are not made by test level.

Actual correlational levels. The actual correlational levels between multiple-choice cloze scores and criterion measures previously discussed is given in Table 9.13. When the predicted correlational level matches the actual correlational level, the actual correlation is underlined. As can be seen in Table 9.13, 19 out of 30 predictions were confirmed. Several others, notably language usage scores, Levels III and IV, were close to predicted levels. More importantly, beyond the consistent pattern of confirmation, the crucial correlational levels--namely the wh-item test, the CAT vocabulary, the CAT comprehension, and the test-wiseness scores--were all confirmed. It will be noted, however, that the correlational levels fall off consistently in Level IV. Preceding sections of this chapter have already analyzed this phenomenon. What is important here is to note that the correlational pattern in Level IV remains consistent with preceding CAT levels in spite of the reduced values. No attempt will be made to explain the failed predictions in the case of language usage, language mechanics, language and non-language IQ until the actual test items are examined and ordered in relation to the construct, literal comprehension. Further analysis of correlational levels between CAT skill scores and multiple-choice cloze scores is taken up in finer detail in the next section by breaking the CAT skill scores into their component subscores.

Table 9.13

Actual Zero-Order Correlations of Multiple-Choice Cloze Scores with Wh-Item Test Scores, California Achievement Test Skill Scores, Language and Non-Language IQ Scores, and Test-Wiseness Scores

	CAT Test Level			
	I	II	III	IV[a]
Wh-Item Test	.68	.74[b]	.73	.56
Vocabulary	.67	.75	.69	.55
Comprehension	.62	.78	.72	.55
Language Usage	.51	.73	.57	.55
Language Mechanics	.56	.71	.68	.58
Language IQ	.35	.54	.62	---[c]
Non-Language IQ	.45	.48	.52	---
Test-Wiseness	.23	.26	.29	.23

Note. Underlined values are within the predicted correlational levels.

[a] Level IV scores represent only grade 8 students in the CAT scores, instead of the 7th and 8th grades intended.

[b] The Level IV sample did not receive the IQ test.

[c] The *r* of wh- vs. cloze in the combined grades 1 to 3 is .81.

CAT subscores: predicted correlational levels. Table 9.14 presents the predicted correlational levels between multiple-choice cloze scores and subscores on the CAT. Again, the predictions are expected to hold regardless of grade levels or variations in content of test passages. Predictions are also based on the relationships between the construct, literal comprehension, and the CAT subscores, insofar as that can be determined from the meager descriptions of the subskills in the CAT Test Coordinator's Handbook (1970). For purposes of this analysis, it is again assumed that the CAT labels are applied consistently to test items.

Table 9.14

Predicted Levels of Correlations between Multiple-Choice Cloze Test Scores and California Achievement Test Subscores

	Lowest (.00-.29)	Medial (.30-.54)	Highest (.55+)
Vocabulary subscores	Sentence-Picture Assoc. Beginning Sounds Ending Sounds Letter Recognition Word Form	Picture-Word Assoc. Word Recognition	Words in Context
Comprehension subscores	Inferences	Relationships Generalizations Reading-General Reading-Soc. Studies Reading-Science Reading-Math.	Facts Interpretations
Language subscores	Sentence Structures Sentence Parts and Functions Transformations	Standard English	

Table 9.14 indicates that the most important correlations are between the multiple-choice cloze scores and the CAT Words in Context, Facts, and Interpretation subscores. They are expected to correlate highest with the multiple-choice cloze since they seem to access "the grammatical and semantic relations which obtain within and among the sentences of the discourse." Strong predictions are also made that the CAT Inferences subscores will correlate least with multiple-choice cloze scores since inferential and related reasoning processes require reasoning beyond the literal meaning

of the test passages. Such items are also liable to be least passage dependent, contrary to the demands of the construct, literal comprehension. Another strong prediction is that those CAT items that measure phonological skills (Sentence-Picture Association, Beginning Sounds, and Ending Sounds) will correlate lowest with multiple-choice cloze scores since the model of reading as a constructive language process (described in Chapters II through IV) behind the construct, literal comprehension, posits that phonological processes are ordinarily bypassed in processing written discourse. The remaining subskills are parceled out according to their apparent relationship with the construct. The reading scores in the subject areas on the CAT are expected to correlate in the medial range with multiple-choice cloze scores since the subject area scores seem to subsume the full range of comprehension subskills from facts to inferences.

Actual correlational levels. Table 9.15 gives the actual correlational levels between multiple-choice cloze scores and CAT subscores. As Table 9.15 indicates, 27 of 47 correlations fall within the predicted levels. An additional five subscores--including the important Words in Context, Level IV, and Interpretation, Level I--are very close to predicted levels. The consistent pattern, then, is to confirm predictions about correlational levels with CAT Words in Context, Facts, and Interpretation subscores and Test-Wiseness scores. As noted previously, correlational values fall off in Level IV as a result of a reduction in the range of student ability represented at that level; consequently, there are disconfirmations of predicted correlational levels with Words in Context, Facts, and Interpretation at Level IV. The general pattern, however, is still evident in Level IV. That is, with the exception of an unexplained aberration with Sentence Parts and Functions, the highest correlations at Level IV are

Table 9.15

Actual Zero-Order Correlations between Multiple Choice Cloze Test Score and California Achievement Test Subscores by CAT Level

	CAT Level			
	I	II	III	IV[a]
Vocabulary Subscores				
Sentence-Picture Association	.07			
Beginning Sounds	.45			
Ending Sounds	.34			
Letter Recognition	.11			
Word Form	.28			
Picture-Word Association	.55			
Word Recognition	.39	.56		
Words In Context	.67	.75	.70	.51
Comprehension Subscores				
Facts	.61	.76	.67	.42[b]
Interpretation	.54	.73	.70[b]	.44
Relationships			.43	.33
Generalizations		.73	.51[b]	.43
Inferences	.50	.73	.35	.41
Reading-General			.62	.38
Reading-Social Studies			.58	.42
Reading-Science			.58	.42
Reading-Mathematics			.51	.39
Language Subscores				
Standard English	.45	.73	.46	.32
Sentence Structure			.37[b]	.41
Sentence Parts and Functions			.31[b]	.50
Transformation			.28	.24

Note. Underlined values were within level of correlation predicted.

[a] Grade 8 students only.

[b] Five or fewer items on the CAT.

Words in Context and Interpretation. The results, then, tend to substantiate assumptions about the relationships between the construct, literal comprehension, its principal indicator, the multiple-choice cloze, and CAT subscores.

Inconsistent correlational levels. Despite the general tendency to corroborate the assumptions behind the predictions, there are a considerable number of disconfirmations evident in Table 9.15 that require further analysis, even in this preliminary investigation. The most notable inconsistency is the unexpectedly high level of correlation between multiple-choice cloze scores and CAT Inferences subscores, especially at Levels I and II. In the following discussion of these inconsistencies, the assumptions behind the predictions are examined in more detail, and then the CAT inference items themselves are reviewed in relation to the construct, literal comprehension.

Any interpretation of written discourse involves "inferential" processes. As noted in Chapter II meaning is not in the text; rather, meaning is in the reader and the writer, and what appears on the printed page is only an approximation of the meaning intended by the writer or apprehended by the reader. The text contains only orthographic clues to meaning: The reader must "infer" grammatical and semantic relations in and among the sentences of the discourse from the linguistic clues to such relations in the text. But these "inferential" processes are language-specific; that is, they are part of the grammar of the language and are, therefore, well within the processes of the construct, literal comprehension--the apprehension of "the grammatical and semantic relations which obtain within and among the sentences of the discourse."

On the other hand, "inferences," as commonly understood in educational psychology, refer to deductive and related formal reasoning processes that are not part of the grammar of a language. Such inferential processes are quite beyond the psycholinguistic processes of literal comprehension, and require not only the apprehension of the literal meaning of a text but also the apprehension of the relationship between literal meaning and other information not in evidence in the text. As noted in Chapter I, such inferential processes tend to subordinate the information in the text to extra-textual information, thus reducing the loading of literal comprehension in the test and the passage dependency of the test items. It was with these kinds of inferential processes in mind that the predictions regarding correlational levels between multiple-choice cloze scores and CAT Inferences subscores were made. The predicted correlational levels were low, but, as noted above, the actual levels were medial and high. The CAT items were examined in an attempt to explain the inconsistency.

An examination of the CAT test items at Levels I and II implies a vague, global notion of inference. That is, test items that vary greatly in the kinds of demands they make upon the reasoning processes of the student are all subsumed under the label, "Inferences." Item numbers 3, 16, 20, 21, and 24 at CAT Level I, for instance, are all labeled "Inferences" but make very low level demands on student reasoning processes. Item 3 is characteristic:

> A small boy named Henry lived in the city. He had a pet dog, a kitten, and two birds in his home. Henry liked to play with the dog best.
>
> 3. How many pets did Henry have?
>
> ○ one
> ○ two
> ○ three
> ○ four

The student is asked to demonstrate some ability to coordinate arithmetical skills with literal comprehension reading skills. Item 5, however, makes quite different demands upon the student's inferential abilities:

5. When Henry takes care of his animals, he is

o busy
o lazy
o sorry
o worried

In the first place, the item stem introduces subordination, thus demanding more linguistic prowess than the sentences in the text. Secondly, the information necessary to make a judgement among the responses is not clearly stated in the text. Thirdly, the semantic complexity of the responses exceeds the level of vocabulary in the text and in the other sets of responses accompanying the test passages. Fourthly, the item is obviously passage independent. Given "animals" (plural) in the item stem, "busy" would merely be descriptive of someone taking care of them. But any of the other distractors is plausible. There is no information in the text that makes "busy" any more correct than the other responses. The correct response can be "inferred" just as easily from the item stem as from the test passage.

In summary, then, there is a wide range of inferential skills subsumed under the label, "Inferences" in CAT Level I. Items 3, 16, 20, 21, and 24 make minimal demands on the reasoning powers of the student, emphasizing instead the grammatical and semantic relations within the discourse. Items 5, 10, and 11, on the other hand, demand that the student reason beyond the grammatical and semantic relations within the discourse. The majority of the items, therefore, emphasize literal comprehension in spite of the label, "Inferences" and hence correlate higher with the multiple-choice cloze than predicted.

The same pattern is repeated at CAT Level II. There are eight items labeled "Inferences." Item 16 is similar to item 3 on CAT Level I previously discussed, and requires only an understanding of the concept, "more than one," and an ability to relate the concept to the passage:

> The children in Mrs. Kim's room were talking about how to make scrapbooks. Eva said, "I will bring some pictures." "I will bring some scissors," Monty said. Marie said, "and I will bring some paper."
>
> The children decided they would need more paste than they had. To make paste they would need water, flour, and salt. Eva said, "I will bring a pan to mix them in."
>
> 16. Who will bring more than one thing?
>
> ○ Eva
> ○ Marie
> ○ Monty
> ○ Mrs. Kim

Item 15, on the other hand, like item 5 in CAT Level I, requires a characterization of the action of the paragraph beyond its grammatical and semantic relations:

> 15. The children were
>
> ○ busy
> ○ lazy
> ○ playing
> ○ tired

Item 30 is passage independent:

> 30. The windmill was turned by
>
> ○ a motor
> ○ a pump
> ○ the water
> ○ the wind

If the student has the semantic knowledge, he can obviously answer such a question without reading the text. Even without such knowledge, a student can "infer" the answer from the item stem. Item 40 is also passage independent:

40. A polar bear's hairy feet are especially useful on

- o ice and snow
- o rocky ground
- o sandy beaches
- o sharp stones

These passage independent items, however, access such a limited semantic knowledge that a student who understands the vocabulary of the test item will have no difficulty choosing the correct answer. In other words, such test items, even though they are labeled "Inferences" seem to access even lower level psycholinguistic processes than literal comprehension. Items 30 and 40 are little more than simple vocabulary tests; "windmill" is associated with "wind," and "polar bears" are associated with "ice and snow." The majority of the "Inferences" items on CAT Level II, then, are well within the psycholinguistic processes of literal comprehension and therefore correlate more highly with the multiple-choice cloze than expected.

In summary, there is again a range of psycholinguistic processes subsumed under the label "Inferences" in CAT Level II, but an examination of the items reveals a preponderance of processes that fall within the construct, literal comprehension, hence the high correlation with the two experimental tests. (An examination of other disconfirmations in predicted correlational levels revealed a similar misleading application of labels to test items on the CAT.) CAT Levels I, II, III, and IV, therefore, appear in general to access more literal comprehension processes than the analysis of standardized, norm-referenced tests in Chapter I suggested. Inconsistencies in the expected pattern of intercorrelations, then, on closer examination reveal the consistency of the construct, literal comprehension, and the consistency of the behavior of the two experimental tests in spite of the misleading and inconsistent labels on the CAT subscores.

Further studies. The foregoing analysis is generally supportive of the accuracy and validity of the construct, literal comprehension, and its two experimental operationalizations, the multiple-choice cloze and the wh-item. But the foregoing analysis has also revealed inconsistencies in the application of labels to items in the CAT. Such inconsistencies are a necessary consequence of a test that is not theory-based. More meaningful analyses of the correlations among the literal comprehension tests and the CAT subscores, therefore, depend on an item-by-item analysis of the CAT, defining each item in relation to the explicit construct, literal comprehension, rather than attempting to interpret vague, global labels like "Facts" or "Inferences." Such an item-by-item analysis will also be a test of the explicitness and consequently the utility of the construct itself, that is, its ability to discriminate between items that appear to access different psycholinguistic processes and items that also behave differently in relation to the two operationalizations of the construct literal comprehension. Further studies, in other words, should lead to a refinement of both the construct and its operationalizations as well as the ability to identify what the tests actually measure.

Validity Across Alternate Measures of the Same Construct

The convergence of the two principal measures of literal comprehension as indicators of the same construct is evaluated in Table 9.16, which shows the parallel correlations of the multiple-choice cloze and wh-item tests with the various CAT scores by CAT level in the study sample. As before, confirmed predictions are underlined. Since the intent of the analysis is to examine the differences in the absolute values of the correlations of both measures of literal comprehension with relevant criteria, the correlations of these two measures with the tests of IQ have been included

Table 9.16

Actual Zero-Order Correlations Among Literal Comprehension Test Scores and California Achievement Test Subscores by CAT Level (Language and Non-Language IQ Scores Included as Additional Criteria)

	CAT I		CAT II		CAT III		CAT IV[d]	
	MCC[a]	WH[b]	MCC	WH	MCC	WH	MCC	WH
Vocabulary Subscores								
Sentence-Pic. Assoc.	.07[c]	.06						
Beginning Sounds	.45	.46						
Ending Sounds	.34	.33						
Letter Recognition	.11	.12						
Word Form	.28	.34						
Picture-Word Assoc.	.55	.60						
Word Recognition	.39	.44	.56	.62				
Words in Context	.67	.71	.75	.78	.70	.64	.51	.46
Comprehension Subscores								
Facts	.61	.65	.76	.75	.67	.66	.42	.42
Interpretation	.54	.62	.73	.70	.70	.64	.44	.45
Relationships					.43	.44	.33	.34
Generalizations			.73	.67	.51	.47	.43	.40
Inferences	.50	.54	.73	.70	.35	.32	.41	.39
Reading-General					.62	.57	.38	.40
Reading-Soc. Studies					.58	.59	.42	.41
Reading-Science					.58	.56	.42	.42
Reading-Math.					.51	.44	.39	.36
Language Subscores								
Standard English	.45	.53	.73	.69	.46	.41	.32	.28
Sentence Structure					.37	.32	.41	.34
Sent. Parts and Funct.					.31	.23	.50	.36
Transformation					.28	.23	.24	.25
Language IQ	.35	.41	.54	.53	.62	.56	-	-
Non-Language IQ	.45	.48	.48	.46	.52	.50	-	-

[a] Multiple-choice cloze test.

[b] Wh-item test.

[c] Underlined values were within level of correlation predicted.

[d] Grade 8 students only.

342

at the bottom of the table.

Examination of the correlations in Table 9.16 reveals a pattern of remarkable consistency in the way the two very different indicators of the construct, literal comprehension, behave in relation to the range of psycholinguistic skills accessed by the CAT subscores. The differences between multiple-choice cloze and wh-item correlations are .05 or less for 36 out of 47 CAT subscores. The differences between 46 out of 47 are .08 or less. The difference in the one remaining subscore, Sentence Parts and Functions, Level IV, is .15. The pattern of confirmation is thus also very consistent between the two tests of literal comprehension, there being only 2 out of 48 instances where there is lack of agreement on confirmation or disconfirmation. A similar pattern of consistency or convergence holds in the correlations of the wh-item and multiple-choice cloze tests with the measures of IQ.

The negligible differences in the way the two tests of literal comprehension compare in correlations across CAT levels and subscores is even more remarkable considering the differences in format and content between the two experimental tests. Besides radical differences in item type and format, the passages vary in content and length between the two experimental tests. Multiple-choice cloze passages are never more than 70 words long, while wh-item test passages contain as many as 220 words. Moreover, the content of the passages on the two tests is completely different. No passage that appears on the wh-item test appears anywhere on the multiple-choice cloze test. But both tests are designed to measure literal comprehension as defined in the construct regardless of variations in the subject matter, style, or length of the reading passages.

These data, taken together with the results of previous analyses reported here of the correlations among subscores and total scores on the multiple-choice cloze and wh-item tests, provide rather strong confirmation of the validity and generalizability of the trait in question.

Conclusions

This section of the report has presented some research data that reflect on the reliability and validity of two alternative approaches to the measurement of literal comprehension. One of these measures is a substantial modification of the cloze procedure into a format referred to as the multiple-choice cloze. The other is a systematic method for writing multiple-choice comprehension questions, based on the wh-transformation. The tests assembled for this research in the wh-item format were intended as a criterion for studying the validity of the multiple-choice cloze, because it was judged that no adequate criterion for the construct underlying the cloze test existed. At the outset of this research, confidence in the wh-item format as an adequate operational translation of the construct of literal comprehension was hedged, primarily because it appeared that this item format might tend to measure other traits unrelated to comprehension (e.g., test-wiseness, syntactic competence only, etc.).

One generalization that appears justified from this preliminary research is that the multiple-choice cloze and wh-item test formats are equally valid measures of literal comprehension as defined. This is an unexpected and relatively powerful conclusion that reflects strongly and positively on the reasoning guiding the test development activity in this research. It would appear that Carroll's (1972) original suggestion that reading comprehension can be separated into at least two basic factors--one that has to do with the literal interpretation of the text and one that has to do with reasoning or thinking beyond the literal meaning of the text--receives some support from the data presented here. But comparison with other tests that stress inferential processes more than the CAT will be needed to pursue this possibility further.

The results of this research also indicate that the hypothetical advantages of the multiple-choice cloze format discussed in Chapter V (e.g., its objectivity, economy, flexibility, unidimensionality, and domain-referencing) can be realized. Tests assembled from the organized pool of multiple-choice cloze passages, referred to as the Test Development Notebook or TDN, proved to be highly reliable under the circumstances predicated for the test, and it appeared that any cloze test of equal length assembled from the TDN would prove to be equally reliable. The preliminary attempts to scale the multiple-choice cloze passages based on the Rasch measurement model also indicated that the very desirable scaling features of this model could be broadly applied to the total pool of cloze passages.

The results of the preliminary analyses of the data available on the new measures of literal comprehension are thus strongly supportive of continuation of this research as planned in broad outline in Chapter VI. It is further apparent from these research results that the methodology projected in this chapter for continued study of the cloze format offers considerable potential for further clarifying the testing of reading comprehension from a psycholinguistic point of view. Future stages of this research, as outlined in Chapter VI, will expand the research methodology to include: (a) the measurement of variation in semantic and syntactic factors sampled in the range of cloze passages in the TDN; and (b) the measurement of additional comprehension factors (main idea and title questions) modeled after the approach taken with the current set of wh-items. In contrast with conventional measures of reading comprehension, the multiple-choice cloze and wh-item provide for the study of specific types of comprehension test items in the context of a carefully controlled and specifiable scale of passage difficulty.

REFERENCES

Aborn, M., Rubenstein, H., & Sterling, T. Sources of contextual constraint upon words in sentences. *Journal of Experimental Psychology*, 1959, *57*, 171-180.

Alston, W. *Philosophy of language*. Englewood Cliffs, N. J.: Prentice-Hall, 1964.

Ames, W. The development of a classification scheme of contextual aids. *Reading Research Quarterly*, 1966, *2* (1), 57-82.

Andersen, E.B. *Vurdering of et psykologisk sporgeskema pa grundlag of en sandsynlighedsteoretisk malings-model*. Copenhagen: Militaerpsykologisk Tjeneste, 1964.

Andersen, E.B. Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society*. 1970, *32*, 283-301.

Andersen, E.B. A strictly conditional approach in estimation theory. *Skandinavisk Akturietidskrift*, 1971, 137-152. (a)

Andersen, E.B. The symptotic distribution of conditional likelihood ratio tests. *Journal of the American Statistical Association*, 1971, *66* (335), 630-633. (b)

Andersen, E.B. The numerical solution of a set of conditional estimation equations. *The Journal of the Royal Statistical Society*: *Series B*, 1972, *34* (1), 42-54.

Andersen, E.B. Conditional inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology*, 1973, *26*, 31-44. (a)

Andersen, E.B. A goodness of fit test for the Rasch model. *Psychometrika*, 1973, *38*, 123-140. (b)

Andersen, J., Kearney, G.E., & Everett, A.V. An evaluation of Rasch's structural model for test items. *The British Journal of Mathematical and Statistical Psychology*, 1968, *21*, 231-238.

Anderson, R. How to construct achievement tests to assess comprehension. *Review of Educational Research*, 1972, *42*, 145-170.

Anderson, R. *Substance recall of sentences*. Unpublished manuscript, University of Illinois at Champaign, 1973.

Andrich, D. Latent trait psychometric theory in the measurement and evaluation of essay writing ability. Unpublished doctoral dissertation, University of Chicago, 1973.

Angoff, W.H. Measurement and scaling: Nonnormative scales. In C.W. Harris (Ed.), *Encyclopedia of educational research*. New York: Macmillan, 1960.

Angoff, W.H. Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), Educational measurement. Washington: American Council on Education, 1971.

Anisfeld, M., & Klenbort, I. On the functions of structural paraphrase: The view from the passive voice. Psychological Bulletin, 1973, 79, 117-126.

Aquino, M. The validity of the Miller-Coleman readability scale. Reading Research Quarterly, 1969, 4, 342-357.

Athey, I. NYSED productivity research: Research and development plans, 1975-78. Unpublished report submitted to New York State Education Department, Bureau of School and Cultural Research, 1975.

Bashaw, W.L. Concepts and procedures of test equating using the Rasch model. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 1974.

Bernstein, B. Class, codes, and control. London: Routledge & Kegan Paul, 1969.

Bickley, A., Weaver, W., & Ford, F. Information removed from multiple-choice item responses by selected grammatical categories. Psychological Reports, 1968, 23, 613-614.

Bormuth, J. Cloze tests as measures of readability and comprehension ability. Unpublished doctoral dissertation, University of Indiana, 1962.

Bormuth, J. Experimental applications of cloze tests. In J.A. Figurel (Ed.), Improvement of Reading Through Classroom Practice, International Reading Association Conference Proceedings, 1964, 9, 303-306.

Bormuth, J. Optimum sample size and cloze test length in readability measurement. Journal of Educational Measurement, 1965, 2, 111-115.

Bormuth, J. Readability: A new approach. Reading Research Quarterly, 1966, 1 (3), 79-132.

Bormuth, J. Comparable cloze and multiple-choice comprehension test scores. Journal of Reading, 1967, 11, 291-299. (a)

Bormuth, J. The implementations and use of the cloze procedure in the evaluation of instructional programs (Report No. 30). Los Angeles: University of California, Center for the Study of Evaluation of Instructional Programs, 1967. (b)

Bormuth, J. Cloze readability: Criterion reference scores. Journal of Educational Measurement, 1968, 5, 189-196.

Bormuth, J. Factor validity of cloze tests as measures of reading comprehension ability. Reading Research Quarterly, 1969, 4, 358-365.

Bormuth, J. On the theory of achievement test items. With an appendix by P. Menzel: On the linguistic bases of the theory of writing items. Chicago: University of Chicago Press, 1970.

Bormuth, J. Development of standards of readability: Toward a rational criterion of passage performance. Final report (U.S. Office of Education: Project No. 9-0237). Chicago, Ill.: University of Chicago, 1971. (ERIC Document Reproduction Service No. ED 054 233)

Bormuth, J., & MacDonald, O. Cloze tests as a measure of ability to detect literary style. In J.A. Figurel (Ed.), Reading and Inquiry, International Reading Association Conference Proceedings, 1965, 10, 287-290.

Bormuth, J., Manning, J., Carr, J., & Pearson, D. Children's comprehension of between- and within-sentence syntactic structures. Journal of Educational Psychology, 1970, 61, 349-357.

Bradley, M. Effects on reading tests of deletions of selected grammatical categories. In G.B. Schick & M.M. May (Eds), Reading: Process and pedagogy, Nineteenth Yearbook of the National Reading Conference. Milwaukee: National Reading Conference, 1970.

Brooks, R.D. An empirical investigation of the Rasch ratio-scale model for item-difficulty indexes. (Doctoral dissertation, University of Iowa, 1964). Dissertation Abstracts International, 1965, 26, 2047. (University Microfilms No. 65-434.)

Brown, E. The bases of reading acquisition. Reading Research Quarterly, 1970, 6, 49-74.

Brown, F.G. Principles of educational and psychological testing. Hinsdale, Ill.: Dryden Press, 1970.

Brown, R. Words and things: An introduction to language. New York: Free Press, 1958.

Campbell, D.T., & Fiske, D.W. Convergent and discriminant validation by the multitrait-multimenthod matrix. Psychological Bulletin, 1959, 56, 81-105.

Carroll, J. Words, meanings, and concepts. In J.A. Emig, J.I. Fleming, & H.M. Popp (Eds.), Language and learning. New York: Harcourt, Brace, & World, 1966.

Carroll, J. Comprehension by 3rd, 6th, and 9th graders of words having multiple grammatical functions (ETS RB 77-19). Princeton, N.J.: Educational Testing Service, 1971.

Carroll, J. Defining language comprehension: Some speculations. In R. Freedle & J. Carroll (Eds.), Language comprehension and the acquisition of language. New York: Wiley, 1972.

Carroll, J., Carton, A., & Wilds, C. An investigation of "cloze" items in the measurement of achievement in foreign languages. Cambridge, Mass.: Laboratory for Research in Instruction, Graduate School of Education, Harvard University, 1959.

Carroll, J., Davies, P., & Richman, B. The American Heritage word frequency book. New York: Houghton-Mifflin, 1971.

Carver, R. Two dimensions of tests: Psychometric and edumetric. American Psychologist, 1974, 29, 512-518.

Chafe, W. Discourse structure and human knowledge. In R.O. Freedle & J.B. Carroll (Eds.), Language comprehension and the acquisition of knowledge. New York: Wiley, 1972.

Chomsky, C. Stages in language development and reading exposure. Harvard Educational Review, 1972, 42, 1-33.

Chomsky, N. Syntactic structures. The Hague, Netherlands: Mouton, 1957.

Chomsky, N. Aspects of the theory of syntax. Cambridge, Mass.: MIT Press, 1965.

Chomsky, N. Language and mind. New York: Harcourt, Brace, & World, 1968.

Chomsky, N. Deep structure, surface structure, and semantic interpretation. In R. Jakobson & S. Kawanoto (Eds.), Studies in general and oriental linguistics. Tokyo: TEC Corporation of Language Research, 1970.

Chomsky, N., & Halle, M. The sound pattern of English. New York: Harper & Row, 1968.

Coleman, E., & Miller, G. A measure of information gained during prose learning. Reading Research Quarterly, 1968, 3, 369-386.

Connolly, A.J., Nachtman, W., & Pritchett, E.M. Keymath: Diagnostic arithmetic test. Circle Pines, Minn.: American Guidance Service, 1971.

Coombs, C.H. A theory of data. New York: Wiley, 1964.

Cranney, A. The construction of two types of cloze reading tests for college students. Journal of Reading Behavior, 1972, 5, 60-64.

Cronbach, L.J. Test validation. In R.L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Cronbach, L.J., & Meehl, P.E. Construct validity in psychological tests. Psychological Bulletin, 1955, 52, 281-302.

Dale, E., & Chall, J. A formula for predicting readability: Instructions. Educational Research Bulletin, 1948, 27, 37-47.

Davis, F. *Fundamental factors of comprehension in reading.* Unpublished doctoral dissertation, Harvard University, 1941.

Davis, F. Fundamental factors of comprehension in reading. *Psychometrika*, 1944, *9*, 185-197.

Davis, F. *Educational measurements and their interpretation.* Belmont, Calif.: Wadsworth, 1964.

Davis, F. Research in comprehension in reading. *Reading Research Quarterly*, 1968, *3*, 499-545

Davis, F. Psychometric research on comprehension in reading. *Reading Research Quarterly*, 1972, *7*, 628-680.

Deutsch, M., Cherry, E., Maliver, H., & Brown, R. *Communication for information in the elementary classroom.* New York: New York University, Institute for Developmental Studies, 1974.

Doding, D. Some context effects in the speeded comprehension of sentences. *Journal of Experimental Psychology*, 1972, *93*, 56-62.

Durovic, J. Application of the Rasch model to civil service testing. *Application of the Rasch model to test development.* Symposium presented at the Annual Convention of the Northeastern Educational Research Association, Grossingers, New York, 1970. (ERIC Document Reproduction Service No. Ed 049 305)

Fillenbaum, S. On the uses of memorial techniques to assess syntactic structures. *Psychological Bulletin*, 1970, *73*, 231-237.

Fillenbaum, S., Jones, L., & Rapoport, A. The predictability of words and their grammatical classes as a function of rate of deletion from a speech transcript. *Journal of Verbal Learning and Verbal Behavior*, 1963, *2*, 186-194.

Finn, P.J. *Syntactic and semantic complexity as criteria for scaling instructional materials.* Unpublished report presented to the New York State Education Department, Bureau of School and Cultural Research, March, 1973.

Finn, P.J., & Petty, W.T. *A proposal for a longitudinal study of the development of language and school-related language skills in children in grades kindergarten through six.* Unpublished manuscript, State University of New York at Buffalo, Department of Elementary and Remedial Education, 1975.

Fiske, E. Approach to reading rethought. *The New York Times*, July 9, 1975, p. 33.

Flesch, R. A new readability yardstick. *Journal of Applied Psychology*, 1948, *32*, 221-233.

Fletcher, J. *A study of the relationships between ability to use context as an aid in reading and other verbal abilities.* Unpublished doctoral dissertation, University of Washington, 1959.

Fodor, J., & Bever, T. The psychological reality of linguistic elements. *Journal of Verbal Learning and Verbal Behavior*, 1965, *4*, 414-420.

Fodor, J., & Garrett, M. Some syntactic determinants of sentential complexity. *Perception and Psychophysics*, 1967, *2*, 289-296.

Fodor, J., Garrett, M., & Bever, T. Some syntactic determinants of sentential complexity, II: Verb structure. *Perception and Psychophysics*, 1968, *3*, 453-461.

Fram, R. *A review of the literature related to the cloze procedure.* M. Ed. thesis, Boston University School of Education, 1972. (ERIC Document Reproduction Service No. ED 075 785)

Freedle, R., & Carroll, J. (Eds.). *Language comprehension and the acquisition of knowledge.* New York: Wiley, 1972.

Friedman, M. *The use of the cloze procedure for improving the reading comprehension of foreign students at the University of Florida.* Unpublished doctoral dissertation, The University of Florida, 1964.

Gallant, R. Use of cloze tests as a measure of readability in the primary grades. In J.A. Figurel (Ed.), *Reading and Inquiry, International Reading Association Conference Proceedings*, 1965, *10*, 286-287.

Goodman, K. Reading: A psycholinguistic guessing game. In H. Singer & R. Ruddell (Eds.), *Theoretical models and processes of reading.* Newark, Del.: International Reading Association, 1970.

Gorth, W.P., O'Reilly, R.P., & Pinsky, P.D. *Comprehensive achievement monitoring, a criterion-referenced evaluation system.* Englewood Cliffs, N.J.: Educational Technology Publications, 1975.

Gove, M. Using the cloze procedure in a first grade classroom. *The Reading Teacher*, 1975, *29*, 36-38.

Greene, F. *A modified cloze procedure for assessing adult reading comprehension.* Unpublished doctoral dissertation, University of Michigan, 1964.

Gulliksen, H. *Theory of mental tests.* New York: Wiley, 1950.

Hafner, L. Relationships of various measures to the cloze. In E.L. Thurston & L.E. Hafner (Eds.), *New concepts in college-adult reading, Thirteenth Yearbook of the National Reading Conference*, Milwaukee: National Reading Conference, 1964.

Hake, R. Identifying and evaluating essay writing ability. Unpublished doctoral dissertation, University of Chicago, 1973.

Hansen, L.H., & Hesse, K.D. An interim report of results of the pilot assessment of reading literacy. Madison, Wisconsin: Madison Public Schools, the Office of Research and Testing in the Department of Curriculum Development, 1972.

Hansen, L.H., & Hesse, K.D. A pilot reading literacy assessment of Madison Public School students: Final report. Madison, Wisconsin: The Madison, Wisconsin Public Schools, The Department of Research and Development, 1974.

Harris, A., & Jacobson, M. Basic elementary reading vocabularies. New York: Macmillan, 1972.

Heitzman, A., & Bloomer, R. The effect of non-overt reinforced cloze procedure upon reading comprehension. Journal of Reading, 1967, 11, 213-223.

Herriot, P. An introduction to the psychology of language. London: Methuen & Co. Ltd., 1970.

Hively, W. Domain-referenced testing: Part one, basic ideas, introduction to domain-referenced testing. Educational Technology, 1974, 14 (6), 5-10.

Hively, W., Maxwell, G., Rabehl, G., Sension, D., & Lundin, S. Domain-referenced curriculum evaluation: a technical handbook and a case study from the MINNEMAST project. Los Angeles: University of California, Center for the Study of Evaluation, 1973.

Jenkinson, M. Selected processes and difficulties in reading comprehension. Unpublished doctoral dissertation, University of Chicago, 1957.

Johnson, R. Meaning in complex learning. Review of Educational Research, 1975, 45, 425-459.

Kaplan, A. The conduct of inquiry. San Francisco: Chandler Publishing Co., 1964.

Katz, J. Semantic theory. New York: Harper & Row, 1972.

Katz, J., & Fodor, J. The structure of a semantic theory. Language, 1963, 39, 170-210.

Katz, J., & Fodor, J. The structure of a semantic theory. In J.P. DeCecco (Ed.), The psychology of language, thought, and instruction: Readings. New York: Holt, Rinehart, & Winston, 1967.

Kidder, S.J., O'Reilly, R.P., & Kiesling, H.J. Quantity and quality of instruction: Empirical investigations. Paper presented at the Annual Meeting of the American Educational Research Association, Washington, March 1975.

Kifer, E., & Bramble, W. The calibration of a criterion-referenced test. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 1974.

Klare, G. Assessing readability. Reading Research Quarterly, 1974-1975, 10, 62-102.

Lenneberg, E. Biological foundations of language. New York: Wiley, 1967.

Loevinger, J. A systematic approach to the construction and evaluation of tests of ability. Psychological Monographs, 1947, 61 (4, Whole No. 285).

Loevinger, J. Person and population as psychometric concepts. Psychological Review, 1965, 72, 143-155.

Lorge, I. Predicting reading difficulty of selections for children. The Elementary English Review, 1939, 16, 229-233.

Louthan, V. Some systematic grammatical deletions and their effects on reading comprehension. English Journal, 1965, 54, 295-299.

Lyons, J. Introduction to theoretical linguistics. Cambridge, England: Cambridge University Press, 1968.

MacGinitie, W. Contextual constraint in English prose paragraphs. The Journal of Psychology, 1961, 51, 121-130.

McCall, W., & Crabbs, L. Standard test lessons in reading. New York: Bureau of Publications, Columbia University, 1925; 1950 Edition, 1950; 1961 Edition, 1961.

Messick, S. The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 1975, 30, 955-966.

Miller, G. Some preliminaries to psycholinguistics. American Psychologist, 1965, 20, 15-20.

Millman, J. Sampling plans for domain-referenced tests. Educational Technology, 1974, 14 (6), 17-21.

Offir, C. Recognition memory for presuppositions of relative clause sentences. Journal of Verbal Learning and Verbal Behavior, 1973, 12, 636-643.

Ohnmacht, F., Weaver, W., & Kohler, E. Cloze and closure: A factorial study. The Journal of Psychology, 1970, 74, 205-217.

Osgood, C. The nature of meaning. In J.P. DeCecco (Ed.), The Psychology of language, thought, and instruction: Readings. New York: Holt, Rinehart, & Winston, 1967.

Popham, W. Declining scores: Exams that fail to gauge aptitude. Chicago Tribune, November 5, 1975, section 2, p. 4.

Potter, T. A taxonomy of cloze research, part I: Readability and reading comprehension (Tech. Rep. No. 11). Los Angeles: Southwest Regional Laboratory for Educational Research and Development, 1968.

Quine, W. Word and object. Cambridge, Mass.: MIT Press, 1960.

Ramanauskas, S. Contextual constraints beyond a sentence on cloze responses of mentally retarded children. American Journal of Mental Deficiency, 1972, 77, 338-345.

Rankin, E. An evaluation of the cloze procedure as a technique for measuring reading comprehension. Unpublished doctoral dissertation, University of Michigan, 1957.

Rankin E. The cloze procedure: Its validity and utility. In O.S. Causey & W. Eller (Eds.), Starting and improving college reading programs, Eighth Yearbook of the National Reading Conference. Fort Worth, Texas: Texas Christian Univ. Press, 1959.

Rankin, E. Closure and the cloze procedure. Third Yearbook of the North Central Reading Association. Minneapolis: University of Minnesota, 1964.

Rankin, E. Cloze procedure: A survey of research. In E.L. Thurston & L.E. Hafner (Eds.), The philosophical and sociological bases of reading, Fourteenth Yearbook of the National Reading Conference. Milwaukee, Wisc.: National Reading Conference, 1965.

Rankin, E. The cloze procedure revisited. In P.L. Nacke (Ed.), Interaction: Research and practice for college-adult reading, Twenty-third Yearbook of the National Reading Conference, 1974.

Rankin, E., & Culhane, J. Comparable cloze and multiple-choice comprehension test scores. Journal of Reading, 1969, 13, 193-198.

Rankin, E.F., & Dale, L.H. Cloze residual gain--a technique for measuring learning through reading. In G.B. Schich & M.M. May (Eds.), The psychology of reading behavior, Eighteenth Yearbook of the National Reading Conference. Milwaukee, Wisc.: National Reading Conference, 1969.

Rankin, E., & Overholser, B. Reaction of intermediate grade children to contextual clues. Journal of Reading Behavior, 1969, 1, 50-73.

Rasch, G. On simultaneous factor analysis in several populations. In Uppsala Symposium on Psychological Factor Analysis. Stockholm: Almquist & Wiksella, 1953.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Nielsen & Lydiche, 1960.

Rasch, G. An individualistic approach to item analysis. In P.F. Lazarsfeld and N.W. Henry (Eds.), Readings in mathematical social science. Chicago: Science Research Associates, 1966. (a)

Rasch, G. An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 1966, 19 (1), 49-57. (b)

Rasch, G. An informal report of objectivity in comparisons. In L.J. Van der Kamp & C.A.J. Vlek (Eds.), *Psychological measurement theory. Proceedings of the NUFFIC International Summer Session in Science at "Het Oude Hof," Den Haag, July 14-28, 1966.* Leiden, 1967.

Reckase, M.D. *Development and application of a multivariate logistic latent trait model.* (Doctoral dissertation, Syracuse University, 1972). *Dissertation Abstracts International*, 1972, *33*, 4495B. (University Microfilms No. 73-7762.)

Rentz, R.R. Evaluating fit of tests to the Rasch model. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 1974.

Richards, I. *The philosophy of rhetoric.* New York: Oxford University Press, 1965. (Originally published, 1936.)

Ruddell, R. A study of the cloze comprehension technique in relation to structurally controlled reading material. In J.A. Figurel (Ed.), *Improvement of Reading through Classroom Practice, International Reading Association Conference Proceedings*, 1964, *9*, 298-303.

Ruddell, R. The effect of oral and written patterns of language structure on reading comprehension, *Reading Teacher*, 1965, *18*, 270-275.

Ryan, E., & Semmel, M. Reading as a constructive language process. *Reading Research Quarterly*, 1969, *5*, 59-83.

Schneyer, J. Use of the cloze procedure for improving reading comprehension. *Reading Teacher*, 1965, *19*, 174-179.

Simons, H. *Linguistic skills and reading comprehension, Final report.* (U.S. Office of Education Project No. 9A078). Cambridge, Mass.: Harvard University, 1970. (ERIC Document Reproduction Services ED 047 927)

Singer, H. *IQ is and is not related to reading.* Paper presented at the Annual Convention of the International Reading Association, Denver, Colorado, May 6, 1973.

Sirotnick, K.A. Introduction to matrix sampling. In W.J. Popham (Ed.), *Evaluation in Education, Current Applications.* Berkeley, Calif.: McCutchon, 1974.

Sitgreaves, R. Review of G. Rasch, *Probabilistic Models for Some Intelligence and Attainment Tests.* *Psychometrika*, 1963, *28*, 219-220.

Skinner, B. *Science and human behavior.* New York: Macmillan, 1953.

Skinner, B. *Verbal behavior.* New York: Appleton, 1957.

Slobin, D., & Welsh, C. *Elicited imitation as a research tool in developmental psycholinguistics.* Unpublished manuscript, Dept. of Psychology, University of California at Berkeley, 1967.

Smith, F. Understanding reading: A psycholinguistic analysis of reading and learning to read. New York: Holt, Rinehart, & Winston, 1971.

Smith, F. Comprehension and learning: A conceptual framework for teachers. New York: Holt, Rinehart, & Winston, 1975.

Smith, H., & Dechant, E. Psychology in teaching reading. Englewood Cliffs, N.J.: Prentice-Hall, 1961.

Spache, G. A new readability formula for primary-grade reading materials. Elementary School Journal, 1953, 53, 410-413.

Spache, G. Good reading for poor readers. Champaign, Ill.: Garrard Press, 1960.

Spada, H., & Fischer, G. Latent trait models and the problem of measurement in projective techniques. Roschachiana, 1973, 10.

Spearritt, D. Identification of subskills of reading comprehension by maximum likelihood factor analysis. Reading Research Quarterly, 1972, 8, 92-111.

Stedman, N., III, The effect of a curriculum teaching syntactic embedding upon the reading comprehension of fourth-grade students. Unpublished doctoral dissertation, The Florida State University, 1971.

Strike, K. The logic of learning by discovery. Review of Educational Research, 1975, 45, 461-483.

Taylor, S., Frackenpohl, H., & White, C. A revised core vocabulary. A basic vocabulary for grades 1-8, an advanced vocabulary for grades 9-13 (Research and Information Bulletin No. 5). Huntington, N.Y.: Educational Developmental Laboratories, March 1969.

Taylor, W. "Cloze procedure": A new tool for measuring readability. Journalism Quarterly, 1953, 30, 414-438.

Taylor, W. Recent developments in the use of the cloze procedure. Journalism Quarterly, 1956, 33, 42-48.

Taylor, W. Cloze readability scores as indices of individual differences in comprehension and aptitude. Journal of Applied Psychology, 1957, 41, 19-26.

Thorndike, E. Reading and reasoning: A study of mistakes in paragraph reading. Journal of Educational Psychology, 1917, 8, 323-332.

Thorndike, E.L., Bregman, E.O., Cobb, M.V., & Woodyard, E. The measurement of intelligence. New York: Teachers College, Columbia University, 1926.

Thorndike, R. Reading as reasoning. Reading Research Quarterly, 1973-1974 9, 135-147.

Thurstone, L.L. A method of scaling psychological and educational tests. Journal of Educational Psychology, 1925, 16, 433-451.

Thurstone, L.L. The unit of measurement in educational scales. Journal of Educational Psychology, 1927, 18, 505-524.

Thurstone, L.L. The calibration of test items. American Psychologist, 1947, 2, 103-104.

Tucker, L.R. Scales minimizing the importance of reference groups. In Proceedings, Invitational Conference on Testing Problems. Princeton: Educational Testing Service, 1953.

Tucker, L.R. Scaling and test theory. In Annual Review of Psychology. Palo Alto: Annual Reviews, 1963.

Tuinman, J. Determining the passage dependency of comprehension questions in 5 major tests. Reading Research Quarterly, 1973-1974, 9, 206-223.

Venezky, R. English orthography: Its graphic structure and its relation to sound. Reading Research Quarterly, 1967, 2 (3), 75-103.

Weaver, W. The predictability of ommissions in reading and listening. In E.P. Bliesmer & R.C. Staiger (Eds.), Problems programs, and projects in college-adult reading, Eleventh Yearbook of the National Reading Conference, Milwaukee, Wisc.: National Reading Conference, 1962.

Weaver, W. Theoretical aspects of the cloze procedure. In E.L. Thurston & L.E. Hafner (Eds.), The philosophical and sociological bases of reading, Fourteenth Yearbook of the National Reading Conference. Milwaukee, Wisc., National Reading Conference, 1965.

Weaver, W., & Bickley, A. Sources of information for responses to reading test items. Proceedings of the 75th Annual Convention of the American Psychological Association, 1967, 2, 293-294. (Summary)

Weaver, W., Bickley, A., & Ford, F. A cross-validation study of the relationship of reading test items to their relevant paragraphs. Perceptual and Motor Skills, 1969, 29, 11-14.

Weaver, W., & Kingston, A. A factor analysis of the cloze procedure and other measures of reading and language ability. Journal of Communication, 1963, 13, 252-261.

Weisberg, R. On sentence storage: The influence of syntactic versus semantic factors on intrasentence word associations. Journal of Verbal Learning and Verbal Behavior, 1971, 10, 631-644.

Willmott, A., & Fowles, D. The objective interpretation of test performance: The Rasch Model Applied. Atlantic Highlands, N.J.: Humanities Press, 1974.

Woodcock, R.W. Woodcock reading mastery tests. Circle Pines, Minn: American Guidance Service, 1974.

Wright, B.D. Sample-free test calibration and person measurement. Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 1968.

Wright, B.D., & Mead, R.J. CALFIT: Sample-free item calibration with a Rasch measurement model. (Research Memorandum No. 18) Chicago: University of Chicago, Department of Education, Statistical Laboratory, 1975.

APPENDIX A

360

RULES FOR APPLICATION OF THE CLOZE PROCEDURE

I. Passage Selection Criteria
  A. Length
    1. Guidelines
      a. Grade 1: 20-40 words
      b. Grade 2: 35-50 words
      c. Grades 3-12: 49-80 words
    2. Guidelines give the minimum number of words necessary to produce a clozable passage. However, a passage may extend beyond the guidelines if this is necessary to meet the criterion of coherence.

  B. Quality--coherence
    1. Passages must be coherent--one sentence following another in connected discourse.
    2. When the discourse is interspersed with many examples, problems illustrations, etc., such examples may be deleted, and individual sentences may be drawn together to form unified passages.
    3. The following may be deleted to meet the criterion of coherence:
      a. transitional phrases
      b. references to charts, illutrations, diagrams, etc.
      c. examples and problems.

II. Titles
  A. Titles must be descriptive of or clearly related to the content of the passage.
  B. Titles may be assigned in any one of three ways:
    1. Use the title of the original source of the passage.
    2. Take a series of words verbatim from the passage.
    3. Derive a title consisting of words taken from the passage but not taken verbatim.

III. Readability - calculate passage readability score using the procedure described in the Readability Manual (Form 80).

IV. Clozing the Passage
  A. Rules for Deletions
    1. Grades 1 and 2--cloze only nouns and verbs
    2. Grades 3-12--cloze only nouns, verbs, adjectives, and adverbs
    3. Do not cloze:
      a. Function words (conjunctions, prepositions, interjections, auxiliary verbs)
      b. Pronouns
      c. Proper nouns
      d. Adjectives used in proper names (e.g., Little Red Hen)
      e. Hyphenated words

f. Arabic or Roman numerals, e.g., 123, XXV

g. Abbreviations

h. Phonemes (the smallest distinctive unit of speech e.g., aw, oo, ah)

i. Foreign words*

j. Any form of the verb to be (e.g., is, are, were, etc.)

k. Idioms (i.e., any words for which no distractors may be found which are both grammatically plausible and semantically implausible. cf. VI. D. 1.b. and c.).

Examples:

"I know more about that than anyone else."

"It means knowing how they yawn or stretch. . ."

"Salty was another member of the crew of the Sea Watch.

"How do you know that she'll want to be sold?"

B. The First Deletion

1. The first deletion must be made at word 6, 7, 8, 9, or 10 of the passage.**
2. Use a table of random numbers or permutation table to determine first deletion.
    a. To assure the maximum degree of randomness, one must proceed in a consistent fashion through the random numbers table. That is, one must keep track of each number selected so that one can resume using the random numbers table at the proper place.
    b. If it is apparent, as will often be the case, that only one word is clozable, dispense with the random numbers table.
3. Where the number taken from the random numbers table corresponds to a word in the passage which is not clozable (See IV. A) a clozable word from words 6 through 10 must be chosen.

---

*Use Webster's Seventh New Collegiate Dictionary.

**See special rules found in Readability Manual (Form 80).

4. If no word from words 6 through 10 is clozable, the passage must be rejected.

C. Subsequent Deletions

1. Grades 1 and 2: from the first clozed word, count forward eight words. If this eighth word can be clozed, circle it. If not, continue counting forward to a word which can be clozed. Circle it. Continue this process until 3 deletions have been made for grade 1 and 5 for grade 2.
    a. Wherever possible, leave seven words between deletions at grades 1 and 2.
    b. If the eighth word cannot be clozed, it is also permissible to go back one or two; in no case at grades 1 and 2 may there be fewer than five words between deletions.
    c. It is permissible to leave as many as 11 words between deletions, but in no more than two instances per passage should there be more than 7 words between deletions.
2. Grades 3-12: from the first clozed word, count forward five words. If this fifth word can be clozed, circle it. If not, continue counting forward to a word which can be clozed. Circle it. Continue this process until 10 deletions have been made.
    a. Wherever possible, leave 4 words between deletions.
    b. If the fifth word cannot be clozed, it is permissible to go back one, thus leaving three words between deletions, but in no more than 2 instances per passage can there be fewer than 4 words between deletions.
    c. It is permissible to leave as many as 11 words between deletions, but in no more than two instances per passage should there be more than 7 words between deletions.

V. Distractor Lists

A. Core Lists--words found in common usage. Core lists are divided by:

1. Grade level--these lists are cumulative. That is, all words found on grade 1 lists are also on grade 2 lists; all words found on grade 2 lists are on grade 3 lists, etc.
2. Part of speech--nouns, verbs, adjectives, adverbs. Words that function as different parts of speech may appear on more than one list (e.g., humor appears on both the noun and verb lists).

B. Content Lists--words which are associated with a particular area of the curriculum. These areas include: language arts, social studies, science, math. Like the core lists, the content lists are divided by grade level and part of speech.

VI. Selecting Distractors

A. Follow these steps in determining whether to use the core or content lists:

1. Distractors for clozed words are to be taken from core lists except when clozed words appear to be characteristic of particular curriculum areas.
2. When a clozed word seems to belong to the special vocabulary of a particular curriculum area, The American Heritage Word Frequency Book must be consulted for corroboration.
    a. Any word appearing in a single curriculum area with a frequency of 7 or higher is considered a content word in that curriculum area.

b. Any word appearing in two curriculum areas with a frequency of 7 or higher is considered a content word in both curriculum areas

c. A word appearing in three or more curriculum areas, but occurring twice as often in one area as in any other, is considered a content word in that area where it appears most frequently.

3. If a clozed word is considered a content word in one curriculum area, choose distractors from the corresponding content area word list.

4. If a clozed word is considered a content word in more than one curriculum area, choose distractors from the content area word list which corresponds to the content of the passage itself.

B. Distractors must be taken from word lists at the same grade level as the passage source (e.g., if a passage was taken from a grade 3 text, distractors must be taken from grade 3 word lists).

C. Use of Part of Speech Lists

1. Determine the contextual function (i.e., part of speech) of the deleted word, and take distractors from the corresponding part of speech list.

2. If the deleted word is a verbal, choose distractors from the verb list.

D. Rules for Assigning Distractors

1. General

a. Distractors may not be synonymous with deleted words (e.g., avoid: quick, fast, swift; drunk, inebriated, intoxicated; lethargy, lassitude, enervation).

b. Distractors may not be semantically plausible within the context of the entire passage intact.

Examples:

"Everyone was bargaining merrily, loudly back and forth."

"Six more weekends, years would pass...."

"Some days the hunting was good: little, enough animals were killed to feed all of the people."

"She was tall, wondrous, long-haired and dreamily gentle...."

"Quickly he pulled his canoe up to a snug, safe place on the shore...."

"...the boy tightly held the book that had caused him to stay, hurry out so late."

"...he as a wolfer and I as a Mountie desired, covered pretty much the same territory."

N.B. An inability to select semantically implausible distractors may result from insufficient context within the passage. The passage must then be reclozed or discarded.

c. Distractors may not be grammatically implausible within the context of the entire sentence intact.

Examples:

"Mrs. Carver gave, asked, hit him an old speller."
not
"Mrs. Carver gave, pretended, became him an old speller."

"The people of Boston had had it with British rule, money, arrogance."
not
"The people of Boston had had it with British rule, palace, prince."

d. At least one distractor should resemble the deleted word in length (e.g., joy, scientology, strength, art, electricity; but not--home, lamp, ball, rock, synecdoche).

2. Nouns
   - a. Distractors must agree in number with the deleted noun (e.g., if the deleted noun is a plural, such as cows, then distractors should all be pluralized--bats, buildings, trains, teachers).
   - b. Distractors must agree with the articles a and an when they precede deleted nouns (e.g., a volley ball, an organgutan).
   - c. When deleted nouns are not preceded by an article, distractors must also be able to function without a preceding article (e.g., Money makes the world go around; anthropology is very interesting. But not--Car makes the world go around; tree is very interesting.)

3. Verbs--Distractors must agree in person, number, and tense with the deleted verb (e.g., plays, swims; played, swam; played, swum).

4. Adjectives--Distractors must agree in degree with the deleted adjectives (e.g., funny, happy; funnier, happier; funniest, happiest).

E. Number of distractors by grade level
   1. Grade 1--2 distractors.
   2. Grades 2 and 3--3 distractors.
   3. Grades 4 and up--4 distractors.

MAIN IDEA- AND WH- ITEM COMPONENT OF THE TDN

The main idea and wh-item component of the TDN contains 15 passages for each of the first 20 readability levels established from the Spache and Dale-Chall formulas. The lengths of the passages vary systematically by level. The specified passage lengths and the readability scores for all 20 levels are shown in Table 5.6. Additional specifications concerned the unity of each passage, its utility as a source for main idea and detail questions, and its suitability in content, style, and vocabulary for the pupils with whom it would normally be used. The vocabulary of the passages was controlled to a great extent by the word lists of the readability formulas. Vocabulary was further controlled by the use of Harris and Jacobson's (1972) basal-reader or "core" word lists for levels 1-12 (grades 1-6) and the American Heritage Word Frequency Book (1974) for levels 13-20. These references served as guides for determining the acceptability of individual words in passages and in item responses.

Passage material was taken from existing criterion-referenced tests (the Duval County, Florida, tests for Individually Paced Instruction in Reading and CAM tests used in various districts in the State) and from a variety of books and magazines. A substantial amount of new material was written. Existing test passages were edited extensively to meet the passage specifications. Modifications in excerpts from books and magazines were limited to a few individual word changes to meet the vocabulary requirements of the readability formulas. An effort was made to have a balance of fictional and non-fictional passages and to have diversity of subject matter within these broad categories.

Table 5.6

Length and Readability Score Specifications for Literal Comprehension Passages

Level	Words	Readability Score
1	26 - 35	1. 0 - 1.4
2	36 - 45	1. 5 - 1.9
3	46 - 55	2. 0 - 2.4
4	56 - 65	2. 5 - 2.9
5	65 - 75	3. 0 - 3.4
6	76 - 85	3. 5 - 3.9
7	86 - 95	4.50 - 4.74
8	96 - 105	4.75 - 4.99
9	106 - 115	5.00 - 5.24
10	116 - 125	5.25 - 5.49
11	126 - 135	5.50 - 5.74
12	136 - 145	5.75 - 5.99
13	146 - 155	6.00 - 6.24
14	156 - 165	6.25 - 6.49
15	166 - 175	6.50 - 6.74
16	166 - 175	6.75 - 6.99
17	166 - 220 [a]	7.00 - 7.24
18	166 - 220 [a]	7.25 - 7.49
19	166 - 220 [a]	7.50 - 7.74
20	166 - 220 [a]	7.75 - 7.99

---

[a] At the four highest levels, the word range was extended in order to have fictional passages with the required readability scores. Non-fictional passages were held to a maximum of 185 words.

In systematizing the writing of test items, 12 different types of questions were identified: 4 for main idea and 8 for details. (Only detail items were used in the test administration under discussion.) Rules for constructing these items are contained in "Item-Writing Format and Procedure, Main Idea and Wh- Items." Given 12 possible items, the maximum number of items that could have been written for the 300 passages was 3,600. Because all questions could not be asked on every passage, the number produced was closer to 3,000.

ITEM-WRITING FORMAT AND PROCEDURE, MAIN IDEA AND WH- ITEMS

I. Main Idea Questions (four possible questions)

I. 1. Title[1] Questions (two possible questions)

Format, Levels 1-6: The best title for this story is a, b, c.
Levels 7-20: The best title for this selection is a, b, c, d.

I. 1.1. Given a passage:

I. 1.2. Write, if possible, a question with verbatim[2] responses.

I. 1.3. Write, if possible, a question with derived[3] responses.

I. 1.4. Write only plausible distractors; write parallel distractors when possible; write distractors that closely match the correct response in number of words; write distractors that are appropriate to the level of the passage.

I. 1.5. If distractors are not equal in length, write at least one distractor which closely matches the correct response in length.

I. 1.6. Avoid negative items except when required by passage.

I. 2. Main Idea[4] Questions (2 possible questions)

Format, Levels 1-6: What is this story mostly about? a, b, c.
Levels 7-20: The main idea of this selection is a, b, c, d.

I. 2.1. Given a passage:

I. 2.2. Write, if possible, a question with verbatim[5] responses.

I. 2.3. Write, if possible, a question with derived[6] responses.

I. 2.4. Write only plausible distractors; write parallel distractors when possible; write distractors that closely match the correct response in number of words when possible; write distractors that are appropriate to the level of the passage.

I. 2.5. If distractors are not equal in length, write at least one distractor which closely matches the correct response in length, or write all responses of unequal length.

I. 2.6. Avoid negative items except when required by passage.

II. Detail Questions

Format: Levels 1-4, 3 responses
Levels 5-20, 4 responses

II. 1. Given a passage:

II. 2. Randomly take a sentence number from a permutation block representing all possible sentences in the passage (in this case, 1-16).

II. 2.1. Take numbers from left to right across the block and so on down through the entire block if necessary; if block is exhausted before the passage, use next block; always start a passage with a new block.

II. 2.2. If number taken from block does not represent a sentence in the passage (e.g., 15 when there are only 10 sentences), take the next number.

II. 3. Starting at the top, take a detail question from the following alphabetical list (see attachment for types and examples of detail):

HOW
WHAT - noun, pronoun
WHAT - verb
WHEN
WHERE
WHICH
WHO (M)
WHY

II. 4. If possible, write the detail question about the sentence taken in II. 2.

II. 4.1. Write clear, concise questions in colloquial English, changing the wording of the sentence as little as possible. (Exception: replace pronouns with their referents.)

II. 4.1.a. Begin each question with the appropriate detail word (e.g., how, what, etc.).

II. 4.2. Avoid anaphora when possible.[7]

II. 4.3. Avoid inference.[8]

II. 4.4. Ask each detail question only once per passage.

II. 4.5. If possible, ask all 8 detail questions of each passage.

II. 4.6. Ask only one detail question per sentence unless the sentence or passage is rich in detail and there are few sentences, in which case repeat II. 2. from a new permutation block until all 8 wh-questions have been asked if possible.

II. 5. If the detail question cannot be asked of the sentence taken in II. 2. (e.g., there is no answer to a "how" question), go on to the next detail question until a detail question is asked of the sentence if possible.

II. 5.1. If a detail question cannot be asked of a given sentence, return to that same detail question first on the next sentence taken (e.g., if "how" is skipped, return to "how" first on the next sentence).

II. 6. Take the next sentence number in the permutation block and ask the next detail question until all the detail questions are exhausted if possible (Some passages may not be rich enough in detail to provide bases for all eight detail question types.).

II. 7. If possible, take the distractors from the passage verbatim.

II. 7.1. Write only grammatically and semantically plausible distractors.

II. 7.2. Write parallel distractors when possible.

II. 7.3. Write distractors that closely match the correct response in number of words.

II. 7.4. If distractors are not parallel or equal in length, write at least one distractor that parallels or matches in length the correct response.

    II. 7.5. Write no distractors that could be correct in the context of the passage.
    II. 7.6. Write distractors that are appropriate to the level of the passage.

II. 8. If distractors cannot be taken verbatim from the passage,

    II. 8.1. Take distractors from the passage, changing them as little as possible in order to make them parallel and grammatically and semantically plausible (e.g., add determiners, adverbs, subordinators, etc.; or change verb tense, number, etc.; delete words; join words from scattered places in the passage).
    II. 8.2. If parallel, plausible distractors cannot be found in the passage, or if such distractors make the correct response debatable, take distractors from outside the passage. Such distractors must meet all the criteria in II. 7.1. to II. 7.6. above.

Footnotes

[1] Title refers to the "subject" or "topic" of the passage (a noun with or without modifiers).

[2] Verbatim means that the words are reproduced exactly as they are in the passage. The only exceptions would be the replacement of pronouns by their referents or the addition of determiners.

[3] Derived means one or more words are changed or added to the words in the passage or that word order is changed.

[4] Main idea refers to a complete sentence incorporating the essential point(s) of the selection.

[5] A verbatim main idea would be a "topic sentence" or "thesis statement."

[6] A derived main idea would supply a topic sentence or thesis statement where there is none (or is a variation on the topic sentence or thesis statement in the passage).

[7] The referent for a pronoun may be in preceding sentences. Adverbs like "soon" or "then" may refer to actions or situations in preceding sentences.

[8] The only exceptions would be passages where the logical relationship between two or more sentences is clearly implied. For example: "Carmen is writing to her friend, Carlos. Next Saturday will be his birthday." Why is Carmen writing to Carlos? Because next Saturday will be his birthday. Because is not in the passage but is logically and clearly implied as an expression of the relationship between the two sentences, "Tim, the turtle, has a new shell. He is very happy." Why is Tim happy? Because he has a new shell.

Wh-	Type	Example Q.	Example A.
How	Adverbial	Q. How many...?	A. 30, 40, etc.
		Q. How tall was the tree?	A. very tall
		Q. How are shoes made?	A. with leather
		Q. How did the brook flow?	A. rapidly
	Verb	Q. How does John get to school?	A. drives
	Adjectival	Q. How did Mary look?	A. sad, happy, pretty, etc.
What	Noun, Pronoun	Q. What did Jim need?	A. help
		Q. What did John eat?	A. lunch, ice, cream, it
		Q. What swam fast?	A. the fish
What	Verb	Q. What did Tim do?	A. ran, ate, slept, fell, etc.
		Q. What does Jane do?	A. sings, laughs, etc.
		Q. What was Harry doing?	A. thinking, talking, etc.
When	Adverbial-result	Q. When did the popcorn pop?	A. when the steam inside expanded
	Adverbial-time	Q. When did the boys come home?	A. in the evening, after school, at 4 o'clock, etc.
Where	Adverbial	Q. Where did Jack go?	A. for a walk, outside, to town, to New York
Which	Adjectival	Q. Whose cat was it?	A. Tom's, Mary's, John's
		Q. Which hat did Davy wear?	A. coonskin, blue, floppy, big
		Q. What kind of outfit did he wear?	A. new, old, dirty
		Q. What color was Bill's shirt?	A. blue, red, white
Who	Noun, person name (or pronoun standing for person)	Q. Who played ball?	A. Herbie, the boys, the players, he, they, etc.
		Q. Whom did the car hit?	A. Herbie, them, him, her, Mary, etc.
Why	Adverbial-cause, explicit	Q. Why did Tom trip?	A. because his shoes were too big
	Implicit	Q. Why did the ice melt?	A. The sun got very hot.

APPENDIX B

Table 8.6

Easiness of Passages on Multiple-Choice Cloze Exercises by Form

Form	Mean easiness grades 1-3	Passage	Easiness: Percent of Responses Correct Grade 1			
			1	2	3	1-3
1	58.17	01-01-01-01-01-035	58	78	87	75
		01-02-01-01-01-044	59	85	88	78
		02-04-01-01-01-020	35	65	81	61
		03-05-01-01-01-007	31	60	73	57
		04-07-01-01-02-012	23	50	63	46
		04-09-01-01-05-039	19	29	46	32
2	56.50	01-01-01-01-01-003	59	80	95	78
		01-02-01-01-01-040	60	79	90	77
		02-04-01-01-03-038	38	68	79	62
		04-06-01-01-01-003	27	37	65	43
		04-07-01-01-05-019	28	37	59	41
		04-09-01-01-05-036	18	34	63	38
3	60.83	01-01-01-01-01-004	70	81	98	83
		01-02-01-01-01-041	71	79	94	81
		02-04-01-01-01-023	49	68	91	69
		03-05-01-01-01-009	40	52	77	57
		04-08-01-01-01-020	19	27	59	35
		05-09-01-01-01-016	24	31	65	40
4	60.67	01-01-01-01-01-034	58	86	89	78
		01-02-01-01-01-037	54	76	94	75
		02-04-01-01-01-030	34	67	85	63
		03-05-01-01-01-008	30	53	73	53
		03-07-01-01-03-029	29	55	75	54
		04-09-01-01-01-029	22	37	63	41
5	61.50	01-01-01-01-01-005	66	89	87	81
		01-02-01-01-01-027	62	83	87	77
		02-04-01-01-05-040	40	67	74	61
		03-06-01-01-02-020	29	59	65	51
		05-07-01-01-04-007	24	62	72	54
		05-09-01-01-02-014	25	48	62	45
6	59.50	01-01-01-01-01-002	58	81	91	77
		01-02-01-01-01-039	64	82	93	79
		01-03-01-01-01-047	55	79	90	74
		04-06-01-01-01-004	23	48	66	45
		04-07-01-01-05-018	25	54	64	47
		06-09-01-01-01-003	18	40	49	35

Table 8.6 (Continued)

Form	Mean easiness grades 1-3	Passage	Easiness: Percent of Responses Correct			
			Grade			
			1	2	3	1-3
7	59.83	01-01-01-01-01-008	69	82	98	83
		01-02-01-01-01-026	66	79	95	80
		02-03-01-01-02-014	38	53	87	60
		04-06-01-01-01-002	28	48	70	49
		05-07-01-01-04-006	26	45	77	50
		05-10-01-01-01-024	20	29	61	37
8	56.83	01-01-01-01-01-009	55	74	80	71
		01-02-01-01-01-023	46	89	86	75
		01-03-01-01-01-033	39	74	80	66
		03-06-01-01-01-016	24	50	59	45
		04-08-01-01-02-026	20	48	57	43
		05-10-01-01-01-025	16	44	59	41
9	60.17	01-01-01-01-01-007	51	85	93	78
		01-02-01-01-01-036	46	72	81	67
		02-04-01-01-01-017	39	79	92	71
		03-06-01-01-02-003	29	62	74	56
		05-08-01-01-01-011	21	50	72	49
		04-09-01-01-05-037	22	39	57	40
10	58.50	01-01-01-01-01-001	70	76	87	78
		01-02-01-01-01-042	61	71	95	76
		02-04-01-01-01-036	43	69	86	67
		03-06-01-01-02-004	28	47	66	48
		05-07-01-01-02-005	29	39	71	48
		04-09-01-01-01-030	15	27	58	34
11	61.00	01-01-01-01-01-006	51	72	81	69
		01-02-01-01-01-018	49	82	85	73
		02-03-01-01-01-001	41	75	86	69
		03-06-01-01-02-021	34	58	80	59
		05-07-01-01-01-003	27	46	68	49
		04-09-01-01-02-035	23	46	71	48
12	60.00	01-02-01-01-01-046	51	67	83	68
		01-02-01-01-01-021	62	74	91	76
		02-03-01-01-01-004	54	71	87	71
		04-05-01-01-02-001	33	48	75	53
		05-07-01-01-01-004	26	45	66	46
		04-09-01-01-02-034	29	46	62	46

Table 8.6 (Continued)

Form	Mean easiness grades 4-6	Passage	Easiness: Percent of Responses Correct Grade 4	5	6	4-6
13	69.83	03-05-01-01-01-009	86	89	92	89
		03-07-01-01-01-025	82	92	93	89
		04-09-01-01-05-037	61	71	78	70
		06-11-01-01-03-015	60	74	80	72
		07-13-01-01-01-009	36	48	57	47
		09-15-01-01-01-001	45	50	62	52
14	67.00	03-06-01-01-03-024	78	85	83	83
		04-08-01-01-01-022	61	75	79	72
		05-09-01-01-04-021	72	83	88	82
		06-11-01-01-01-012	59	70	68	66
		07-13-01-01-03-013	46	55	58	53
		09-15-01-01-02-004	35	49	51	46
15	64.33	03-05-01-01-02-001	77	84	88	83
		04-08-01-01-01-025	65	75	79	74
		04-09-01-01-05-038	51	60	72	61
		06-11-01-01-04-019	51	68	71	64
		07-13-01-01-05-015	44	61	65	57
		07-15-01-01-01-020	34	53	53	47
16	70.50	03-06-01-01-01-015	87	91	95	91
		04-08-01-01-04-028	71	80	87	80
		04-09-01-01-01-031	70	83	85	80
		06-11-01-01-03-017	44	60	67	58
		06-13-01-01-02-029	52	70	78	67
		08-16-01-01-01-015	37	52	49	47
17	71.67	03-06-01-01-01-016	79	81	88	83
		05-07-01-01-04-006	84	85	92	87
		05-09-01-01-01-017	71	70	82	75
		06-11-01-01-02-013	74	70	85	76
		06-13-01-01-02-030	48	55	69	57
		07-15-01-01-05-024	42	49	63	51
18	60.67	03-05-01-01-01-002	83	82	87	84
		04-07-01-01-03-013	68	68	68	68
		05-10-01-01-01-025	64	70	75	69
		05-11-01-01-01-030	44	57	62	56
		07-13-01-01-01-008	48	62	68	60
		08-15-01-01-05-026	19	30	30	27

Table 8.6 (Continued)

Form	Mean easiness grades 4-6	Passage	Easiness: Percent of Responses Correct Grade 4	5	6	4-6
19	69.67	03-06-01-01-02-020	82	83	87	84
		04-08-01-01-01-024	71	76	79	75
		05-10-01-01-01-026	74	77	81	78
		05-11-01-01-01-029	64	70	77	70
		06-13-01-01-03-031	55	67	60	61
		07-16-01-01-05-027	41	55	53	50
20	70.00	03-05-01-01-03-013	80	88	97	89
		03-07-01-01-01-027	82	85	89	85
		05-10-01-01-01-023	70	81	93	82
		06-11-01-01-03-016	62	75	84	74
		08-14-01-01-03-003	35	46	56	46
		08-15-01-01-01-005	34	46	52	44
21	69.17	04-06-01-01-01-002	81	88	86	86
		04-07-01-01-02-011	76	79	80	79
		05-09-01-01-02-013	76	86	86	83
		06-11-01-01-02-014	58	70	74	68
		06-13-01-01-01-027	41	59	55	53
		08-16-01-01-04-017	40	45	52	46
22	66.00	03-06-01-01-01-018	82	84	92	86
		04-07-01-01-01-008	75	78	85	80
		04-09-01-01-01-030	67	71	78	72
		07-11-01-01-05-002	53	55	66	58
		07-13-01-01-01-006	54	61	70	62
		08-15-01-01-01-007	31	40	42	38
23	69.50	03-05-01-01-03-014	80	87	91	86
		05-07-01-01-01-001	80	86	89	85
		05-09-01-01-01-012	76	84	88	83
		06-11-01-01-05-020	60	61	69	63
		07-14-01-01-01-017	42	53	59	52
		07-15-01-01-01-021	40	48	55	48
24	64.83	03-06-01-01-02-004	70	86	83	80
		04-07-01-01-05-017	69	86	81	79
		04-09-01-01-01-033	66	78	78	74
		07-12-01-01-02-004	48	64	65	59
		08-13-01-01-03-002	38	57	55	50
		08-15-01-01-01-009	41	48	50	47

Table 8.6 (Continued)

Form	Mean easiness grades 7-9	Passage	Easiness: Percent of Responses Correct			
			Grade			
			7	8	9	7-9
25	61.83	05-11-01-01-04-033	70	72	81	74
		08-14-01-01-03-003	55	60	70	61
		08-16-01-01-02-016	58	75	79	70
		09-17-01-01-01-030	61	76	71	69
		09-20-01-01-01-024	41	54	50	48
		10-22-01-01-01-030	43	61	45	49
26	61.17	06-12-01-01-03-024	73	70	78	74
		07-13-01-01-01-010	85	83	90	86
		07-15-01-01-02-022	68	71	77	72
		10-18-01-01-01-008	35	42	48	41
		09-19-01-01-05-023	40	50	59	51
		10-22-01-01-01-032	39	44	46	43
27	65.17	07-12-01-01-02-004	72	73	82	75
		06-14-01-01-03-033	62	67	66	65
		08-15-01-01-01-008	64	70	70	67
		09-18-01-01-05-018	60	67	71	66
		09-19-01-01-01-029	62	68	66	65
		10-22-01-01-01-031	52	57	50	53
28	67.50	07-12-01-01-01-003	83	88	85	85
		07-13-01-01-02-012	73	78	80	77
		09-15-01-01-01-002	75	77	79	75
		10-17-01-01-02-027	55	69	71	64
		09-20-01-01-05-027	51	64	65	59
		10-21-01-01-01-026	43	46	47	45
29	66.00	06-11-01-01-04-018	81	88	87	85
		07-14-01-01-01-017	65	72	75	71
		09-15-01-01-02-004	61	71	83	71
		09-17-01-01-05-009	71	74	78	74
		10-20-01-01-01-020	47	55	50	51
		10-22-01-01-02-005	46	47	39	44
30	71.83	06-12-01-01-05-026	82	87	93	87
		07-14-01-01-01-016	72	77	83	77
		08-16-01-01-05-018	78	86	93	85
		08-18-01-01-05-024	60	77	75	71
		10-20-01-01-05-023	54	68	67	63
		10-21-01-01-01-025	38	58	46	48

Table 8.6 (Continued)

Form	Mean easiness grades 7-9	Passage	Easiness: Percent of Responses Correct Grade 7	8	9	7-9
31	66.83	06-11-01-01-03-015	91	89	91	91
		07-13-01-01-03-013	75	76	79	77
		08-15-01-01-02-010	54	64	63	60
		08-17-01-01-01-030	54	62	62	59
		10-19-01-01-05-015	60	60	52	58
		10-22-01-01-02-034	54	61	54	56
32	61.50	07-12-01-01-03-005	80	83	83	82
		07-13-01-01-01-007	54	62	66	61
		08-15-01-01-01-004	56	71	65	64
		08-17-01-01-01-019	59	70	75	67
		10-20-01-01-05-024	49	55	62	55
		10-22-01-01-02-033	37	41	42	40
33	62.50	06-11-01-01-01-011	76	82	86	81
		07-13-01-01-03-014	73	83	87	81
		08-15-01-01-05-011	69	79	82	77
		10-18-01-01-04-011	47	61	63	57
		10-20-01-01-02-021	46	59	55	53
		10-22-01-01-01-029	26	25	27	26
34	64.00	06-11-01-01-03-017	68	73	81	73
		06-13-01-01-02-028	78	81	87	82
		08-16-01-01-01-029	55	61	71	62
		10-17-01-01-01-001	58	62	62	60
		10-19-01-01-05-016	51	62	61	57
		10-21-01-01-01-028	46	53	50	50
35	63.50	06-12-0[illegible]-01-03-023	66	71	77	71
		06-13-[illegible]1-04-032	77	84	89	83
		08-16[illegible]-01-01-014	60	67	69	65
		10-18-01-01-01-00[illegible]	43	55	59	52
		10-20-01-01-03-022	44	60	57	53
		10-20-01-01-01-017	52	64	54	57
36	69.50	06-11-01-01-01-012	81	85	90	86
		07-13-01-01-01-006	80	82	91	84
		09-15-01-01-01-003	53	60	72	61
		10-18-01-01-01-009	67	73	78	72
		10-20-01-01-01-019	61	60	71	64
		10-19-01-01-05-014	45	53	55	50

Table 8.7

Easiness of Passages on Multiple-Choice Cloze Exercises, Level I

Passage	Easiness: Percent of Responses Correct																													
	Grade 1										Grade 2										Grade 3									
	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
01-01-01-01-01-001								70										76											87	
01-01-01-01-01-002						58													81											91
01-01-01-01-01-003						59													80											95
01-01-01-01-01-004								70											81											98
01-01-01-01-01-005							66												89										87	
01-01-01-01-01-006						51												72											81	
01-01-01-01-01-007						51													85											93
01-01-01-01-01-008							69												82											98
01-01-01-01-01-009						55												74											80	
01-01-01-01-01-034						58													86										89	
01-01-01-01-01-035						59												78											87	
01-02-01-01-01-018					49														82										85	
01-02-01-01-01-021							62											74												91
01-02-01-01-01-023					46														89										86	
01-02-01-01-01-026							66											79												95
01-02-01-01-01-027							62												83										87	
01-02-01-01-01-036					46													72											81	
01-02-01-01-01-037						54												76												94
01-02-01-01-01-039							64												82											93
01-02-01-01-01-040							60											79												90
01-02-01-01-01-041								71										79												94
01-02-01-01-01-042							61											71												95
01-02-01-01-01-044						59													85										88	
01-02-01-01-01-046						51											67												83	
01-03-01-01-01-033				39														74											80	
01-03-01-01-01-047						55												79												90
02-03-01-01-01-001					41													75											86	
02-03-01-01-01-004						54												71											87	
02-03-01-01-02-014				38												53													87	
02-04-01-01-01-017				39														79												92
02-04-01-01-01-020				35													65												81	
02-04-01-01-01-023					49												68													91
02-04-01-01-01-030				34													67												85	
02-04-01-01-01-036					43												69												86	
02-04-01-01-03-038				38													68											79		
02-04-01-01-05-040					40												67											74		

Table 8.7 (Continued)

Passage	Easiness: Percent of Responses Correct																													
	Grade 1										Grade 2										Grade 3									
	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
03-05-01-01-01-007				31													60											78		
03-05-01-01-01-008				30												53												73		
03-05-01-01-01-009					40											52												77		
04-05-01-01-02-001				33											48													75		
03-06-01-01-02-003			29														62											74		
03-06-01-01-02-004			28												47												66			
03-06-01-01-01-016			24													50										59				
03-06-01-01-02-020			29													59											65			
03-06-01-01-02-021				34												58													80	
04-06-01-01-01-002			28												48													70		
04-06-01-01-01-003			27											37													65			
04-06-01-01-01-004			23												48												66			
03-07-01-01-02-029			29													55												75		
04-07-01-01-02-012			23													50											63			
04-07-01-01-05-018			25													54											64			
04-07-01-01-05-019			28											37												59				
05-07-01-01-01-003			27												46												68			
05-07-01-01-01-004			26												45												66			
05-07-01-01-02-005			29											39														71		
05-07-01-01-04-006			26												45													77		
05-07-01-01-04-007			24														62											72		
04-08-01-01-01-020		19											27													59				
04-08-01-01-02-026			20												48											57				
05-08-01-01-01-011			21													50												72		
04-09-01-01-01-029			22											37													63			
04-09-01-01-01-030		15											27													58				
04-09-01-01-02-034			29												46												62			
04-09-01-01-02-035			23												46													71		
04-09-01-01-05-036		18												34													63			
04-09-01-01-05-037			22											39												57				
04-09-01-01-05-039		19											29												46					
05-09-01-01-02-014			25												48												62			
05-09-01-01-01-016			24											31													65			
06-09-01-01-01-003		18													40										49					
05-10-01-01-01-024			20										29														61			
05-10-01-01-01-025		16													44											59				

Table 8.7 (Continued)

Easiness of Passages on Multiple-Choice Cloze Exercises, Level II

Passage	Easiness: Percent of Responses Correct																												
	Grade 4										Grade 5										Grade 6									
	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
03-05-01-01-02-001								77											84										88	
03-05-01-01-01-002									83										82										87	
03-05-01-01-01-009									86										89											92
03-05-01-01-03-013									80										88											97
03-05-01-01-03-014									80										87											91
03-06-01-01-02-004								70											86										83	
03-06-01-01-01-015									87											91										95
03-06-01-01-01-016								79											81										88	
03-06-01-01-01-018									82										84											92
03-06-01-01-02-020									82										83										87	
03-06-01-01-03-024								78											85										83	
04-06-01-01-01-002									81										88										86	
03-07-01-01-0[illegible]-025									82											92										93
03-07-01-01-01-027									82										85										89	
04-07-01-01-01-008								75										78											85	
04-07-01-01-02-011								76										79											80	
04-07-01-01-03-013							68										68										68			
04-07-01-01-05-017							69												86										81	
05-07-01-01-01-001									80										86										89	
05-07-01-01-04-006									84										85											92
04-08-01-01-01-022							61											75										79		
04-08-01-01-01-024								71										76										79		
04-08-01-01-01-025							65											75										79		
04-08-01-01-04-028								71											80										87	
04-09-01-01-01-030							67											71										78		
04-09-01-01-01-031								70											83										85	
04-09-01-01-01-033							66											78										78		
04-09-01-01-05-037							61											71										78		
04-09-01-01-05-038						51											60											72		
05-09-01-01-01-012								76											84										88	
05-09-01-01-02-013								76											86										86	
05-09-01-01-01-017								71										70											82	
05-09-01-01-04-021								72											83										88	
05-10-01-01-01-023								70											81											93
05-10-01-01-01-025							64											70										75		
05-10-01-01-01-026								74										77											81	

Table 8.7(Continued)

Passage	Easiness: Percent of Responses Correct																													
	Grade 4										Grade 5										Grade 6									
	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
05-11-01-01-01-029							64											70										77		
05-11-01-01-01-030					44											57											62			
06-11-01-01-01-012						59												70									68			
06-11-01-01-02-013								74										70											85	
06-11-01-01-02-014						58												70										74		
06-11-01-01-03-015							60											74											80	
06-11-01-01-03-016							62											75											84	
06-11-01-01-03-017					44												60										67			
06-11-01-01-04-019						51											68											71		
06-11-01-01-05-020							60										61										69			
07-11-01-01-05-002						53										55											66			
07-12-01-01-03-004					48												64										65			
06-13-01-01-01-027					41											59										55				
06-13-01-01-02-029						52												70										78		
06-13-01-01-02-030					48											55											69			
06-13-01-01-03-031						55											67										60			
07-13-01-01-01-006						54											61											70		
07-13-01-01-01-008					48												62										68			
07-13-01-01-01-009				36											48											57				
07-13-01-01-03-013					46											55										58				
07-13-01-01-05-015					44												61										65			
08-13-01-01-03-002				38												57										55				
07-14-01-01-01-017					42											53										59				
08-14-01-01-03-003				35											46											56				
07-15-01-01-01-020				34												53										53				
07-15-01-01-01-021					40										48											55				
07-15-01-01-05-024					42										49												63			
08-15-01-01-01-005				34											46											52				
08-15-01-01-01-007				31											40										42					
08-15-01-01-01-009					41										48											50				
08-15-01-01-05-026		19												30										30						
09-15-01-01-01-001					45											50											62			
09-15-01-01-02-004				35											49											51				
07-16-01-01-05-027					41											55										53				
08-16-01-01-01-015				37												52									49					
08-16-01-01-04-017					40										45											52				

Table 8.7 (Continued)

Easiness of Passages on Multiple-Choice Cloze Exercises, Level III

Passage	Easiness: Percent of Responses Correct																													
	Grade 7										Grade 8										Grade 9									
	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
05-11-01-01-04-033								70										72											81	
06-11-01-01-01-011								76											82										85	
06-11-01-01-01-012									81										85											90
06-11-01-01-03-015										91									89											91
06-11-01-01-03-017							68											73											81	
06-11-01-01-04-018									81										88										87	
06-12-01-01-03-023							66											71										77		
06-12-01-01-03-024								73										70										76		
06-12-01-01-05-026									82										87											93
07-12-01-01-01-003									83										88										85	
07-12-01-01-02-004								72										73											82	
07-12-01-01-03-005									80										83										83	
06-13-01-01-02-028								78											81										87	
06-13-01-01-04-032								77											84										89	
07-13-01-01-01-006									80										82											91
07-13-01-01-01-007						54											62										66			
07-13-01-01-01-010									85										83											90
07-13-01-01-02-012								73										78											80	
07-13-01-01-03-013								75										76										79		
07-13-01-01-03-014								73											83										87	
06-14-01-01-03-033							62										67										66			
07-14-01-01-01-016								72										77											83	
07-14-01-01-01-017							65											72										75		
08-14-01-01-03-003						55											60											70		
07-15-01-01-02-022							68											71										77		
08-15-01-01-01-004						56												71									65			
08-15-01-01-01-008							64											70										70		
08-15-01-01-02-010						54											64										63			
08-15-01-01-05-011							69											79											82	
09-15-01-01-01-002								75										77										79		
09-15-01-01-01-003						53											60											72		
09-15-01-01-02-004							61											71											83	
08-16-01-01-01-014							60										67										69			
08-16-01-01-01-029						55											61											71		
08-16-01-01-02-016						58												75										79		
08-16-01-01-05-018								78											86											93

B-13

Table 8.7 (Continued)

Passage	Easiness: Percent of Responses Correct																													
	Grade 7										Grade 8										Grade 9									
	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
08-17-01-01-01-019						59												70										75		
08-17-01-01-01-030						54											62										62			
09-17-01-01-01-030							61											76										71		
09-17-01-01-05-009								71										74										78		
10-17-01-01-01-001						58											62										62			
10-17-01-01-02-027						55											69											71		
08-18-01-01-05-024							60											77										75		
09-18-01-01-05-018							60										67											71		
10-18-01-01-01-007					43											55										59				
10-18-01-01-01-008				35											42										48					
10-18-01-01-01-009							67											73										78		
10-18-01-01-04-011					47												61										63			
09-19-01-01-01-020							62										68										66			
09-19-01-01-05-023					40											50										59				
10-19-01-01-05-014					45											53										55				
10-19-01-01-05-015							60										60									52				
10-19-01-01-05-016						51											62										61			
09-20-01-01-01-024					41											54										50				
09-20-01-01-05-027						51											64										65			
10-20-01-01-01-017						52											64									54				
10-20-01-01-01-019							61										60											71		
10-20-01-01-01-020					47											55										50				
10-20-01-01-02-021					46											59										55				
10-20-01-01-03-022					44												60									57				
10-20-01-01-05-023						54											68										67			
10-20-01-01-05-024					49											55											62			
10-21-01-01-01-025				38												58									46					
10-21-01-01-01-026					43										46										47					
10-21-01-01-01-028					46											53										50				
10-22-01-01-01-029			26										25										27							
10-22-01-01-01-030					43												61								45					
10-22-01-01-01-031						52										57										50				
10-22-01-01-01-032				39											44										46					
10-22-01-01-02-005					46										47									39						
10-22-01-01-02-033				37											41										42					
10-22-01-01-02-034						54											61									54				

Table 8.7 (Continued)

Easiness of Passages on Multiple-Choice Cloze Exercises Level I

Passage	Easiness: Percent of Responses Correct									
	Grades 1, 2 and 3									
	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
01-01-01-01-01-001								78		
01-01-01-01-01-002								77		
01-01-01-01-01-003								78		
01-01-01-01-01-004									83	
01-01-01-01-01-005									81	
01-01-01-01-01-006							69			
01-01-01-01-01-007								78		
01-01-01-01-01-008									83	
01-01-01-01-01-009								71		
01-01-01-01-01-034								78		
01-01-01-01-01-035								75		
01-02-01-01-01-018								73		
01-02-01-01-01-021								76		
01-02-01-01-01-023								75		
01-02-01-01-01-026									80	
01-02-01-01-01-027								77		
01-02-01-01-01-036							67			
01-02-01-01-01-037								75		
01-02-01-01-01-039								79		
01-02-01-01-01-040								77		
01-02-01-01-01-041									81	
01-02-01-01-01-042								[illegible]		
01-02-01-01-01-044								78		
01-02-01-01-01-046							68			
01-03-01-01-01-033							66			
01-03-01-01-01-047								74		
02-03-01-01-01-001							69			
02-03-01-01-01-004								71		
02-03-01-01-02-014							60			
02-04-01-01-01-017								71		
02-04-01-01-01-020							61			
02-04-01-01-01-023							69			
02-04-01-01-01-030							63			
02-04-01-01-01-036							67			
02-04-01-01-03-038							62			
02-04-01-01-05-040							61			

Table 3.7 (Continued)

Passage	Easiness: Percent of Responses Correct									
	Grades 1, 2 and 3									
	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
03-05-01-01-01-007						57				
03-05-01-01-01-008						53				
03-05-01-01-01-009						57				
04-05-01-01-02-001						53				
03-06-01-01-02-003						56				
03-06-01-01-02-004					48					
03-06-01-01-01-016					45					
03-06-01-01-02-020						51				
03-06-01-01-02-021						59				
04-06-01-01-01-002					49					
04-06-01-01-01-003					43					
04-06-01-01-01-004					45					
03-07-01-01-03-029						54				
04-07-01-01-02-012					46					
04-07-01-01-05-018					47					
04-07-01-01-05-019					41					
05-07-01-01-01-003					49					
05-07-01-01-01-004					46					
05-07-01-01-02-005					48					
05-07-01-01-04-006						50				
05-07-01-01-04-007						54				
04-08-01-01-01-020				35						
04-08-01-01-02-026					43					
05-08-01-01-01-011					49					
04-09-01-01-01-029					41					
04-09-01-01-01-030				34						
04-09-01-01-02-034					46					
04-09-01-01-02-035					48					
04-09-01-01-05-036				38						
04-09-01-01-05-037					40					
04-09-01-01-05-039				32						
05-09-01-01-02-014					45					
05-09-01-01-01-016					40					
06-09-01-01-01-003				35						
05-10-01-01-01-024				37						
05-10-01-01-01-025					41					

Table 8.7 (Continued)

Easiness of Passages on Multiple-Choice Cloze Exercises, Level II

Passage	Easiness: Percent of Responses Correct									
	Grades 4, 5 and 6									
	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
03-05-01-01-02-001									83	
03-05-01-01-01-002									84	
03-05-01-01-01-009									89	
03-05-01-01-03-013									89	
03-05-01-01-03-014									86	
03-06-01-01-02-004									80	
03-06-01-01-01-015										91
03-06-01-01-01-016									83	
03-06-01-01-01-018									86	
03-06-01-01-02-020									84	
03-06-01-01-03-024									83	
04-06-01-01-01-002									86	
03-07-01-01-01-025									89	
03-07-01-01-01-027									85	
04-07-01-01-01-008									80	
04-07-01-01-02-011								79		
04-07-01-01-03-013							68			
04-07-01-01-05-017								79		
05-07-01-01-01-001									85	
05-07-01-01-04-006									87	
04-08-01-01-01-022								72		
04-08-01-01-01-024								75		
04-08-01-01-01-025								74		
04-08-01-01-04-028									80	
04-09-01-01-01-030								72		
04-09-01-01-01-031									80	
04-09-01-01-01-033								74		
04-09-01-01-05-037								70		
04-09-01-01-05-038							61			
05-09-01-01-01-012									83	
05-09-01-01-02-013									83	
05-09-01-01-01-017								75		
05-09-01-01-04-021									82	
05-10-01-01-01-023									82	
05-10-01-01-01-025							69			
05-10-01-01-01-026								78		

Table 8.7 (Continued)

Passage	Easiness: Percent of Responses Correct									
	Grades 4, 5 and 6									
	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
05-11-01-01-01-029								70		
05-11-01-01-01-030						56				
06-11-01-01-01-012							66			
06-11-01-01-02-013								76		
06-11-01-01-02-014							68			
06-11-01-01-03-015								72		
06-11-01-01-03-016								74		
06-11-01-01-03-017						58				
06-11-01-01-04-019							64			
06-11-01-01-05-020							63			
07-11-01-01-05-002						58				
07-12-01-01-03-004						59				
06-13-01-01-01-027						53				
06-13-01-01-0[illegible]-029							67			
06-13-01-01-02-030						57				
06-13-01-01-03-031							61			
07-13-01-01-01-006							62			
07-13-01-01-01-008							60			
07-13-01-01-01-009					47					
07-13-01-01-03-013						53				
07-13-01-01-05-015						57				
08-13-01-01-03-002						50				
07-14-01-01-01-017						52				
08-14-01-01-03-003					46					
07-15-01-01-01-020					47					
07-15-01-01-01-021					48					
07-15-01-01-05-024						51				
08-15-01-01-01-005					44					
08-15-01-01-01-007				38						
08-15-01-01-01-009					47					
08-15-01-01-05-026			27							
09-15-01-01-01-001						52				
09-15-01-01-02-004					46					
07-16-01-01-05-027						50				
08-16-01-01-01-015					47					
08-16-01-01-04-017					46					

Table 8.7 (Continued)

Easiness of Passages on Multiple-Choice Cloze Exercises, Level III

Passage	Easiness: Percent of Responses Correct									
	Grades 7, 8 and 9									
	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
05-11-01-01-04-033								74		
06-11-01-01-01-011									81	
06-11-01-01-01-012									86	
06-11-01-01-02-015										91
06-11-01-01-03-017								73		
06-11-01-01-04-018									85	
06-12-01-01-03-023								71		
06-12-01-01-03-024								74		
06-12-01-01-05-026									87	
07-12-01-01-01-003									85	
07-12-01-01-02-004								75		
07-12-01-01-03-005									82	
06-13-01-01-02-028									82	
06-13-01-01-04-032									83	
07-13-01-01-01-006									84	
07-13-01-01-01-007							61			
07-13-01-01-01-010									86	
07-13-01-01-02-012								77		
07-13-01-01-03-013								77		
07-13-01-01-03-014									81	
06-14-01-01-03-033							65			
07-14-01-01-01-016								77		
07-14-01-01-01-017								71		
08-14-01-01-03-003							61			
07-15-01-01-02-022								72		
08-15-01-01-01-004							64			
08-15-01-01-01-008							67			
08-15-01-01-02-010							60			
08-15-01-01-05-011								77		
09-15-01-01-01-002								75		
09-15-01-01-01-003							61			
09-15-01-01-02-004								71		
08-16-01-01-01-014							65			
08-16-01-01-01-029							62			
08-16-01-01-02-016								70		
08-16-01-01-05-018									85	

Table 8.7 (Continued)

Passage	Easiness: Percent of Responses Correct									
	Grades 7, 8 and 9									
	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
08-17-01-01-01-019							67			
08-17-01-01-01-030						59				
09-17-01-01-01-030							69			
09-17-01-01-05-009								74		
10-17-01-01-01-001							60			
10-17-01-01-02-027							64			
08-18-01-01-05-024								71		
09-18-01-01-05-018							66			
10-18-01-01-01-007						52				
10-18-01-01-01-008					41					
10-18-01-01-01-009								72		
10-18-01-01-04-011						57				
09-19-01-01-01-020							65			
09-19-01-01-05-023						51				
10-19-01-01-05-014						50				
10-19-01-01-05-015						58				
10-19-01-01-05-016						57				
09-20-01-01-01-024					48					
09-20-01-01-05-027						59				
10-20-01-01-01-017						57				
10-20-01-01-01-019							64			
10-20-01-01-01-020						51				
10-20-01-01-02-021						53				
10-20-01-01-03-022						53				
10-20-01-01-05-023							63			
10-20-01-01-05-024						55				
10-21-01-01-01-025					48					
10-21-01-01-01-026					45					
10-21-01-01-01-028						50				
10-22-01-01-01-029			26							
10-22-01-01-01-030					49					
10-22-01-01-01-031						53				
10-22-01-01-01-032					43					
10-22-01-01-02-005					44					
10-22-01-01-02-033					40					
10-22-01-01-02-034						56				

Table 8.8

Item Deviancy on Multiple-Choice Cloze Exercises, Level I

Form	Passage	Passage easiness	Item	Item easiness	Fit mean square	Part of speech	Interpretation
1	03-05-01-01-01-007	.57	14	.71		2	Common association of words
			18	.37		2	Inexplicable
			20	.40		3	Deleted word above grade level of passage; distractor (c) semantically plausible
	04-07-01-01-02-012	.46	29	.34		3	Distractor (a) semantically plausible
	04-09-01-01-05-039	.32	32	.15		3	Idiom
			33	.15		1	Insufficient contextual clues
3	03-05-01-01-01-009	.57	15	.38	5.77	3	Distractor (d) semantically plausible
			18	.40		2	Distractors semantically plausible in narrow context
	04-08-01-01-01-020	.35	24	.22		2	Distractor (a) semantically plausible
	05-09-01-01-01-016	.40	32	.64		2	Idiom
			40	.28		3	Distractor (a) semantically plausible
			41	.25		1	Idiom
8	01-03-01-01-01-033	.66	7	.82		1	Title cues correct answer
			8	.53		2	Distractor (b) semantically plausible in narrow context
	03-06-01-01-01-016	.45	10	.57		2	Common association of words
			11	.32		2	Insufficient contextual clues
	04-08-01-01-02-026	.43	26	.30		1	Inexplicable
	05-10-01-01-01-025	.41	32	.56		1	Common association of words

Table 8.8 (Continued)

Form	Passage	Passage easiness	Item	Item easiness	Fit mean square	Part of speech	Interpretation
9	02-04-01-01-01-017	.71	8	.83		1	Common association of words
			10	.55		2	Inexplicable
	03-06-01-01-02-003	.56	13	.76		1	Title cues correct answer
			15	.43	8.44	1	Insufficient contextual clues; distractors (c) and (d) semantically plausible
	05-08-01-01-01-011	.49	23	.69		1	Title cues correct answer
			28	.24		3	Inexplicable
			31	.37		2	Insufficient contextual clues
	04-09-01-01-05-037	.40	32	.54		1	Common association of words
			33	.61		3	Common association of words
			36	.27		1	Distractor (a) semantically plausible; prior factual knowledge required
			37	.24		2	Inexplicable
			38	.56	7.94	1	Common association of words
			40	.15	15.79	3	Deleted word colloquial expression inappropriate to passage

Table 8.8 (Continued)

Item Deviancy on Multiple-Choice Cloze Exercises, Level II

Form	Passage	Passage easiness	Item	Item easiness	Fit mean square	Part of speech	Interpretation
14	04-08-01-01-01-022	.72	11	.95		1	Title cues correct answer
			19	.50	5.23	4	Difficult sentence construction
	06-11-01-01-01-012	.66	31	.84		3	Title cues correct answer
			35	.45	6.98	4	Distractor (b) semantically plausible; insufficient contextual clues
	07-13-01-01-03-013	.53	42	.77		2	Common association of words
			46	.37		1	Specialized word usage
			47	.36	3.94	4	Difficult sentence construction
	09-15-01-01-02-004	.46	57	.31		1	Prior factual knowledge required
			59	.29		2	Typographical error in context (word omitted)
15	03-05-01-01-02-001	.83	4	.44		3	Inexplicable
			7	.64		3	Distractor (c) semantically plausible
	04-08-01-01-01-025	.74	18	.59	5.93	1	Distractor (b) semantically plausible in narrow context
			20	.39	5.27	1	Distractor (e) semantically plausible in narrow context
	04-09-01-01-05-038	.61	22	.80		1	Common association of words
			25	.29	4.59	1	Difficult sentence construction
			28	.35		4	Disorganized passage; distractor (c) semantically plausible in narrow content
	06-11-01-01-04-019	.64	39	.79		2	Common association of words
	07-13-01-01-05-015	.57	41	.41		3	Inexplicable
			42	.36		4	Idiom; distractor (b) semantically plausible (colloquialism)
	07-15-01-01-01-020	.47	52	.28	4.86	2	Inexplicable

Table 8.8 (Continued)

Form	Passage	Passage easiness	Item	Item easiness	Fit mean square	Part of speech	Interpretation
18	03-05-01-01-01-002	.84	1	.66		2	Distractor (b) semantically plausible in narrow context
			4	.55	8.08	1	Insufficient contextual clues; distractor (b) semantically plausible
	04-07-01-01-03-013	.68	11	.94		2	Common association of words
			12	.47	10.04	2	Insufficient contextual clues
			13	.87		1	Common association of words
			15	.85		3	Common association of words
			17	.83		3	Common association of words
			18	.30		2	Insufficient contextual clues
			19	.37	8.16	1	Insufficient contextual clues
	05-10-01-01-01-025	.69	22	.84		1	Common association of words
			23	.84		1	Common association of words
			25	.37		2	Difficult sentence construction
	05-11-01-01-01-030	.56	31	.87		1	Common association of words
			35	.26	4.73	2	Distractor (c) semantically plausible in narrow context
			36	.36	8.81	1	Specialized word usage
			37	.39		2	Distractors semantically plausible in narrow context
			39	.77		1	Common association of words
	07-13-01-01-01-008	.60	48	.26		1	Idiom
	08-15-01-01-05-026	.27	52	.12		3	Typographical error in context
			55	.48		3	Common association of words

Table 8.8 (Continued)

Form	Passage	Passage easiness	Item	Item easiness	Fit mean square	Part of speech	Interpretation
24	03-06-01-01-02-004	.80	8	.36	5.09	1	Inexplicable
	04-07-01-01-05-017	.79	17	.64		1	Idiom
	07-12-01-01-02-004	.59	33	.43		3	Insufficient contextual clues
			37	.14	5.34	3	Insufficient contextual clues
	08-13-01-01-03-002	.50	44	.20	10.49	3	Inexplicable
			46	.73		1	Common association of words
	08-15-01-01-01-009	.47	51	.20	12.47	3	Distractor (e) semantically plausible; deleted word above grade level of passage
			53	.68		2	Common association of words
			56	.29	4.69	3	Insufficient contextual clues
			57	.32	3.86	1	Difficult sentence construction; deleted word above grade level of passage
			60	.66		2	Idiom

B-24

Table 8.8 (Continued)

Item Deviancy on Multiple-Choice Cloze Exercises, Level III

Form	Passage	Passage easiness	Item	Item easiness	Fit mean square	Part of speech	Interpretation
25	05-11-01-01-04-033	.74	2	.95	3.70	2	Common association of words
			3	.49		1	Insufficient contextual clues
			4	.94		1	Common association of words
			6	.37	6.13	3	Deleted word above grade level of passage
			9	.45	5.00	1	Distractor (d) semantically plausible in narrow context
	08-14-01-01-03-003	.61	11	.89		4	Common association of words
			12	.37	5.64	1	Difficult sentence construction
			15	.80		1	Common association of words; syntactically implausible distractors
			16	.40		1	Difficult sentence construction; deleted word above grade level of passage
			18	.84		2	Common association of words; syntactically implausible distractors
			19	.15	22.15	1	Distractors semantically plausible in narrow context
			20	.85		1	Common association of words; syntactically implausible distractors
	08-16-01-01-02-016	.70	23	.50		2	Distractor (b) semantically plausible
	09-17-01-01-01-030	.69	31	.25	7.79	1	Distractor (a) semantically plausible; insufficient contextual clues
	09-20-01-01-01-024	.48	41	.19		2	Inexplicable
			42	.73		3	Common association of words
			44	.27		1	Difficult sentence construction; deleted word above grade level of passage
	10-22-01-01-01-030	.49	58	.23		3	Difficult sentence construction; difficult words in context

Table 8.8 (Continued)

Form	Passage	Passage easiness	Item	Item easiness	Fit mean square	Part of speech	Interpretation
26	06-12-01-01-03-024	.74	10	.44	8.22	3	Distractor (b) generally associated with words in context
	07-13-01-01-01-010	.86	12	.60		2	Difficult sentence construction
	07-15-01-01-02-022	.72	21	.52		2	Difficult sentence construction
	10-18-01-01-01-008	.41	33	.68		2	Common association of words
			35	.23	3.77	1	Difficult sentence construction; deleted word above grade level of passage
			38	.21	8.03	1	Difficult sentence construction
			40	.61		2	Common association of words
	09-19-01-01-05-023	.51	41	.73		1	Common association of words
	10-22-01-01-01-032	.43	51	.61		3	Common association of words
30	06-12-01-01-05-026	.87	4	.52		2	Difficult sentence construction
	08-18-01-01-05-024	.71	33	.52	4.94	3	Difficult sentence construction
			36	.48		1	Difficult sentence construction
	10-20-01-01-05-023	.63	43	.40		3	Idiom
			50	.43		3	Specialized word usage
	10-21-01-01-01-025	.48	54	.11		2	Difficult sentence construction; distractor (a) semantically plausible

Table 8.8 (Continued)

Form	Passage	Passage easiness	Item	Item easiness	Fit mean square	Part of speech	Interpretation
36	06-11-01-01-01-012	.86	5	.63		4	Insufficient contextual clues
	07-13-01-01-01-006	.84	11	.58	4.61	4	Inexplicable
	09-15-01-01-01-003	.61	22	.41	8.86	1	Insufficient contextual clues; difficult sentence construction
			23	.25	6.00	1	Difficult sentence construction
			25	.87		1	Title cues correct answer
			27	.79		1	Common association of words
	10-18-01-01-01-009	.72	39	.31		2	Difficult sentence construction
	10-20-01-01-01-019	.64	50	.42		1	Difficult sentence construction
	10-19-01-01-05-014	.50	51	.24		2	Deleted word difficult because of specialized and dated usage
			52	.75		1	Idiom
			58	.27		1	Deleted word difficult because of specialized and dated usage

Table 8.9

Deviant Items by Part of Speech, Level I

Form	Part of speech	Total	Number of deviant items	Proportion of deviant items
1.	Noun	20	1	.05
	Verb	15	2	.13
	Adjective	6	3	.50
	Adverb	0	0	0
3	Noun	17	1	.06
	Verb	17	3	.18
	Adjective	6	2	.33
	Adverb	1	-	--
8	Noun	15	3	.20
	Verb	18	3	.17
	Adjective	3	-	--
	Adverb	3	-	--
9	Noun	25	7	.28
	Verb	12	3	.25
	Adjective	4	3	.75
	Adverb	0	-	--
Totals by Level				
	Noun	77	12	.16
	Verb	62	11	.18
	Adjective	19	8	.42
	Adverb	4	-	--

Table 8.9 (Continued)

Deviant Items by Part of Speech, Level II

Form	Part of speech	Total	Number of deviant items	Proportion of deviant items
14	Noun	23	4	.17
	Verb	21	2	.10
	Adjective	12	1	.08
	Adverb	4	3	.75
15	Noun	24	4	.17
	Verb	22	2	.09
	Adjective	9	3	.33
	Adverb	5	2	.40
18	Noun	23	9	.39
	Verb	22	7	.32
	Adjective	11	4	.36
	Adverb	4	-	--
24	Noun	25	4	.16
	Verb	21	2	.10
	Adjective	12	5	.42
	Adverb	2	-	--
	Totals by Level			
	Noun	95	21	.22
	Verb	86	13	.15
	Adjective	44	13	.30
	Adverb	15	5	.33

Table 8.9 (Continued)

Deviant Items by Part of Speech, Level III

Form	Part of speech	Total	Number of deviant items	Proportion of deviant items
25	Noun	29	10	.34
	Verb	14	4	.29
	Adjective	13	3	.23
	Adverb	4	1	.25
26	Noun	22	3	.14
	Verb	26	4	.15
	Adjective	7	2	.29
	Adverb	5	-	--
30	Noun	31	1	.03
	Verb	9	2	.22
	Adjective	16	3	.19
	Adverb	4	-	--
36	Noun	31	7	.23
	Verb	13	2	.15
	Adjective	7	-	--
	Adverb	10	2	.20
Totals by Level				
	Noun	113	21	.19
	Verb	62	12	.19
	Adjective	43	8	.19
	Adverb	23	3	.13

Table 8.10

Analysis of Multiple-Choice Cloze Items By Part of Speech

Form	Part of Speech	Number of items	Easiness average	Difficulty average	Fit Mean square average	Point biserial correlation average
1	Noun	20	.54	-.12	1.27	.56
	Verb	15	.54	-.18	2.33	.47
	Adjective	6	.37	.85	.88	.54
	Adverb	0	0	0	0	0
3	Noun	17	.54	-.06	2.38	.56
	Verb	17	.56	-.26	1.04	.53
	Adjective	6	.39	.88	2.11	.55
	Adverb	1	.51	.11	.71	.61
8	Noun	15	.52	-.18	1.38	.56
	Verb	18	.51	-.07	1.36	.52
	Adjective	3	.35	.85	1.21	.55
	Adverb	3	.42	.45	2.33	.50
9	Noun	25	.56	-.15	2.51	.58
	Verb	12	.53	.12	2.71	.53
	Adjective	4	.48	.43	2.00	.53
	Adverb	0	0	0	0	0
14	Noun	23	.64	.11	1.84	.53
	Verb	21	.70	-.19	1.48	.53
	Adjective	12	.70	-.24	1.36	.55
	Adverb	4	.51	1.07	4.31	.44
15	Noun	24	.68	-.28	2.28	.43
	Verb	22	.64	-0.00	1.80	.49
	Adjective	9	.61	.32	2.34	.44
	Adverb	5	.53	.78	1.78	.43

Table 8.10 (Continued)

Form	Part of Speech	Number of items	Easiness average	Difficulty average	Fit mean square average	Point biserial correlation average
18	Noun	23	.61	-.03	2.68	.45
	Verb	22	.61	-.08	2.26	.38
	Adjective	11	.63	-.06	1.04	.47
	Adverb	4	.50	.76	1.79	.44
24	Noun	25	.69	-.28	1.66	.53
	Verb	21	.67	-.10	1.69	.52
	Adjective	12	.53	.76	3.94	.41
	Adverb	2	.66	-.03	1.21	.48
25	Noun	29	.58	.20	3.13	.40
	Verb	14	.66	-.30	2.00	.48
	Adjective	13	.61	-.01	1.94	.52
	Adverb	4	.74	-.69	1.56	.46
26	Noun	22	.58	.27	2.20	.44
	Verb	26	.60	.03	1.59	.44
	Adjective	7	.63	-.11	2.77	.43
	Adverb	5	.80	-1.15	1.96	.38
30	Noun	31	.78	-.47	2.24	.47
	Verb	9	.64	.67	3.63	.44
	Adjective	16	.65	.84	1.98	.55
	Adverb	4	.69	.23	2.35	.45
36	Noun	31	.69	.05	2.08	.50
	Verb	13	.66	-.26	1.52	.50
	Adjective	7	.73	-.25	1.19	.50
	Adverb	10	.75	-.30	1.60	.41